2013-09-04

Establishing Measurement Invariance of Thin Ideal Internalization and Body Dissatisfaction Across Studies: An Integrative Data Analysis

Kat Tumblin Green
Brigham Young University - Provo

Establishing Measurement Invariance of Thin Ideal Internalization and Body Dissatisfaction Across Studies: An Integrative Data Analysis

Kat Tumblin Green

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Scott A. Baldwin, Chair
Joseph A. Olsen
Jared S. Warren
Mikle South
Dawson W. Hedges

Department of Psychology

Brigham Young University

September 2013

ABSTRACT

Establishing Measurement Invariance of Thin Ideal Internalization and Body Dissatisfaction Across Studies: An Integrative Data Analysis

Kat Tumblin Green
Department of Psychology, BYU
Doctor of Philosophy

With increased data sharing and research collaboration options available through modern technology, there is an increased need to find more advanced techniques to analyze data across multiple studies. A systematic method of pooling participant-level versus study-level data would be particularly valuable as it would allow for more complex statistical analyses, broader assessment of constructs, and a cost effective way to examine new questions and replicate previous findings. One notable difficulty in pooling raw data in the behavioral sciences is the heterogeneity in methodologies and consequent need to establish measurement invariance. The present study explores the feasibility of using Integrative Data Analysis (IDA) to combine 10 heterogeneous eating disorder prevention data sets and establish measurement invariance across the constructs of thin ideal internalization and body dissatisfaction. Using standard multiple groups factor analysis and likelihood-ratio tests to examine differential item functioning, separate one-factor models were established for the three measures used across studies. Partial measurement invariance was established for all measures. Implications for future IDA studies based on this process are discussed, particularly regarding the clinical impact of measurement invariance.

Keywords: Integrative Data Analysis, mega-analysis, measurement invariance

ACKNOWLEDGEMENTS

I would like to thank Scott Baldwin for his help and encouragement throughout this project and for his mentorship throughout my program. I also want to thank Joe Olsen for his willingness to consult and answer questions through the challenges of this project. I am also grateful to Jared Warren and Mikle South for their roles in emphasizing the role of assessment in research and real-world settings, as well as their support and mentorship in the clinical child track. I would also like to thank Dawson Hedges for his willingness to help and consult on this project. I am also very grateful for the support of Eric Stice and Carolyn Becker, whose willingness to share their data with us made this project possible.

I would also like to thank Colby Green for his unwavering support throughout this project. He has been alongside me each step of the way, listening attentively and patiently as I shared my successes and failures in writing statistical code and offering help and support whenever I needed it. I would also like to thank the rest of my family in supporting me in my program and asking about my dissertation, even when they knew what it was about.

Table of Contents

Establishing Measurement Invariance of Thin Ideal Internalization and Body Dissatisfaction Across Studies: An Integrative Data Analysis

Estimates of the prevalence of Bulimia Nervosa among female adolescents and young adults range between 1-4% (Hudson, Hiripi, Pope, & Kessler, 2007; Keski-Rahkonen et al., 2009; Pyle, Neuman, Halvorson, & Mitchell, 1991). Estimates of subclinical bulimia nervosa or bulimic symptoms suggest that almost half of female adolescents and young adults experience significant weight concerns and almost 10% engage in recurrent compensatory purging behaviors and dietary restraint (Ackard, Fulkerson, & Neumark-Sztainer, 2007; Berg, Frazier, & Sherr, 2009; Fairburn et al., 2007; Touchette et al., 2011). Patterns of disordered eating are often long-lasting and research suggests that individuals with eating disordered behaviors and significant weight concerns also experience a higher rate of comorbid emotional problems (Fairburn et al., 2007; Keski-Rahkonen et al., 2009; Touchette et al., 2011). However, despite the significant distress experienced by those with disordered eating, few individuals seek or receive treatment (Hudson et al., 2007; Keski-Rahkonen et al., 2009). In light of the high prevalence of disordered eating, as well as the low proportion of distressed individuals that seek treatment, there is a need for effective prevention interventions for eating disorder symptoms.

Traditional prevention efforts focus on education about symptoms, risks, and consequences of disordered eating, healthy weight management or resisting social pressure. However, these didactic approaches have limited effectiveness (Stice, Shaw, & Marti, 2007). Researchers have suggested three possibilities as to why psychoeducation may not be optimal (Stice, Chase, Stormer, & Appel, 2001; Stice, Mazotti, Weibel, & Agras, 2000). First, information alone may not be enough to change behavior. Second, psychoeducation tends to be delivered to a broad population of individuals rather than to those at an increased risk for disordered eating. Third, many psychoeducational approaches lack a well-developed theory (Stice & Shaw, 2004).

In an effort to address these deficits, Stice and colleagues developed an intervention based on the dual-pathway model of bulimia (Stice, Ziemba, Margolis, & Flick, 1996). This model suggests that sociocultural pressure to be thin promotes internalization of the thin ideal, or the degree to which an individual adopts the societal belief that thinness is attractive. Theoretically, this increased internalization of the thin ideal leads to increases in body dissatisfaction, or the negative evaluation of one's weight or figure. In turn, body dissatisfaction leads to bulimic symptoms through two pathways: dietary restraint and negative affect (Stice, Nemeroff, & Shaw, 1996; Stice & Shaw, 2002; Stice, Ziemba et al., 1996). Based on this model, Stice et al. (2000) developed an intervention targeting thin ideal internalization because it occurs relatively early in the proposed development of bulimia in the dual pathway model and also because it is easier to manipulate than other risk factors, such as body mass or societal pressure. Drawing on principles of cognitive dissonance theory (Festinger, 1957), which suggests that when individuals are presented with inconsistent beliefs, they will change their behavior to try to restore consistency, Stice and colleagues proposed a dissonance-based intervention intended to induce women who had internalized the thin ideal to voluntarily develop an argument against it. They hypothesized that reductions of thin ideal internalization would decrease body dissatisfaction, dietary restraint, negative affect and bulimic symptoms. They also suggested that targeting women with significant concerns about their weight and who are already at risk for disordered eating would be more effective than interventions with a broader scope.

After a preliminary trial by Stice et al. in 2000, there have been trials in six independent labs testing the effectiveness of this dissonance intervention with promising results (e.g., Becker et al., 2010; Green, Scott, Diyankova, Gasser, & Pederson, 2005; Matusek, Wendt, & Wiseman, 2004; Mitchell, Mazzeo, Rausch, & Cooke, 2007; Roehrig, Thompson, Brannick, & van den Berg, 2006; Stice, Rohde, Gau & Shaw, 2009; see Stice, Shaw et al., 2008 for a review). Evidence suggests that the dissonance intervention tends to show better outcomes than waitlist control, assessment only, media advocacy, yoga and healthy weight management groups (Stice, Shaw, et al., 2008). Specifically, the dissonance intervention has shown superior reductions in thin ideal internalization, body dissatisfaction, dietary restraint, negative affect and overall bulimic symptoms (Becker, Bull, Schaumberg, Cauble, & Franco, 2008; Becker, Smith, & Ciao, 2005, 2006; Becker et al., 2010; McMillan, Stice, & Rohde, 2011; Stice, Chase, Stormer, & Appel, 2001; Stice, Marti, Shaw, & Jaconis, 2010; Stice, Marti, Spoor, Presnell, & Shaw, 2008; Stice et al., 2000; Stice et al., 2009; Stice, Shaw, et al., 2008; Stice, Shaw, Burton, & Wade, 2006; Stice, Trost, & Chase, 2003). A recent analysis demonstrated that reductions in the thin ideal lead to decreases in eating disorder risk factors and bulimic symptoms (Stice, Presnell, Gau & Shaw, 2007).

Although various researchers have examined this intervention, a large portion of the available data has been collected in 10 studies conducted by Stice and colleagues (2000, 2001, 2003, 2006, 2009; McMillan et al., 2011) and Becker et al. (2005, 2006, 2008, 2010). All 10 studies conducted by these researchers have tested the same basic principles of the dissonance intervention, but in different formats, across different age ranges and with different populations. All 10 studies have also shown the same general trends in their data, but the reported effect sizes for relationships between variables differ across studies. A valuable next step in extending our understanding and increasing the impact of this dissonance intervention is to meaningfully synthesize findings across these 10 heterogeneous studies. There are a number of ways to combine data across studies, each with unique challenges.

Meta-analysis is a common method for synthesizing results across studies. Meta-analysis uses summary statistics from each study to estimate effect sizes across studies. Findings from several recent meta-analyses in the eating disorder literature have shown positive outcomes of dissonance-based interventions (e.g., Stice, Shaw et al., 2008; Stice, Shaw et al., 2007). In addition to examining overall treatment effects, these meta-analyses have also examined moderators of treatment effect across studies and found that several factors such as older age and multisession formats may improve outcome (Stice, Shaw et al., 2007). However, the use of summary statistics markedly reduces the information available, which in turn limits the moderators that can be examined and the statistical analyses possible (Berlin, Santanna, Schmid, Szczech, & Feldman, 2002; Cooper & Patall, 2009). For example, in Stice, Shaw & Marti (2007), moderators had to be dichotomously coded, where age was examined looking at studies with a mean age greater or less than 15, and the number of sessions was coded as one or more than one.

Researchers have also used a technique called "mega-analysis" to aggregate data across studies. The definition and application of mega-analysis has considerable variability across studies, however. One common use of mega-analysis is aggregating meta-analyses, essentially becoming a meta-analysis of meta-analyses, with the same limitations of standard meta-analyses above (e.g., Cialdini & Fultz, 1990; Freeman & Strayer, 1996). More recently, more mega-analyses have pooled individual patient-level data in an attempt to address some of the limitations of meta-analysis. However, there has not been any systematic way to ensure that measurement properties were the same in all studies, which is crucial to meaningfully interpreting pooled data. Mega-analyses of pooled participant-level data have ranged from disregarding measurement properties to the beginnings of more rigorous but still unsystematic methods in studies that acknowledge the potential for measurement differences and attempt to adjust cut-off scores for included participants accordingly (e.g., de Maat, et al., 2008; DeRubeis, Gelfand, Zang, & Simons, 1999; Fournier et al., 2010; Lambert, Abrams, Sutton & Jones, 2002; Seretti, Cusin, Rausch, Bondy, & Smeraldi, 2006; Sternberg, Baradaran, Abbott, Lamb, & Guterman, 2006).

Integrative Data Analysis

Integrative Data Analysis (IDA) is a systematic method of combining data from multiple studies. IDA allows participant-level data from a modest number of studies to be pooled and analyzed within a single data set (Bauer & Hussong, 2009; Curran & Hussong, 2009; Curran et al., 2008). The studies combined in an IDA must be similar in content and purpose, but they can differ in samples, measures and methodology. There are a number of benefits of using IDA over meta-analysis, when participant-level data is available. These benefits are detailed in a 2009 special issue of *Psychological Methods* (volume 14), and a summary is provided here.

Benefits of IDA

IDA can be used to examine new hypotheses not possible in single studies and with greater flexibility than meta-analysis. This may include questions about overall outcomes across the data sets, as well as moderators within the data such as age, risk-category or intervention format (Cooper & Patall, 2009; Stewart & Clarke, 1995). Evaluating participant-level data often identifies patterns within subsets of the data that are not observed when analyzing only summary statistics (Berlin et al., 2002; Cooper & Patall, 2009). Using IDA to analyze these 10 dissonance-based intervention data sets would allow us to more precisely examine differences in effectiveness of the intervention over a broader developmental range, as ages across studies range from early adolescence to young adulthood. We could also examine include symptom severity, treatment setting and treatment implementation (i.e., duration or treatment administrator). Meta-analysis would allow us to examine these questions on a more limited level, but the additional flexibility that comes with working with raw data means that IDA can provide valuable information about the intervention that can be used to improve and refine the intervention development and implementation.

Another benefit of using IDA is the ability to expand the assessment of constructs and generalizability of measures. The first steps in IDA involve establishing measurement invariance across measures of similar constructs, or assuring that measures assessed the intended construct in a similar way within and across studies. One limitation of most single studies is that the assessment of the construct is limited to the chosen measures and the measurement properties are only evaluated within the populations of individual studies. The use of pooled participant-level data allows researchers to broaden the assessment of different constructs by using multiple, heterogeneous measures, as well as the generalizability of the measures through the use of a much larger, more heterogeneous sample than is present in any single study (Curran & Hussong, 2009). The use of item-level data in establishing measurement invariance also allows researchers to examine moderators of assessment, in addition to moderators of treatment outcomes. For example, researchers could examine the effect of age or other moderators on how individuals respond to assessment items, in addition to the effect on overall treatment effects.

IDA offers a number of other valuable advantages over using group-level summary statistics. For example, having all of the participant-level data allows for the use of more recent and complex multivariate statistical models, increases the statistical power of analyses of intervention effects and mediation/moderation effects, and provides more precise and accurate estimates of variance and effect sizes than can be obtained using meta-analysis (Cooper & Patall, 2009; Lambert, Sutton, Abrams, & Jones, 2002; Stewart & Clarke, 1995). In addition, IDA may be a useful technique for studying low base-rate behaviors. By combining individual samples with low occurrences of a behavior, the base-rate will remain the same in the pooled sample but the number of cases of the behavior increases, allowing for more precise examination of infrequent behaviors (Curran & Hussong, 2009). IDA also allows for an assessment of the replicability/generalizability of findings from single studies without conducting new and costly individual trials. In sum, although IDA is time-intensive (relative to study-level analyses) and more difficult in fields that are heterogeneous in methodology, the development and wider-use of this technique will result in a cost-effective method of using existing data to examine new questions, promote collaboration and shared resources between researchers and increase the quality and effectiveness of research and clinical application across disciplines.

Previous Studies using IDA

There are a number of studies demonstrating the benefits of IDA (e.g., Berlin et al., 2002; Curran et al., 2008; DeRubeis et al., 1999). The majority of IDA studies conducted in the behavioral sciences have used three longitudinal data sets examining alcoholism (Bauer & Hussong, 2009; Curran et al., 2008; Hussong, Bauer, et al., 2008; Hussong, Cai, et al., 2008; Hussong, Flora, Curran, Chassin, & Zucker, et al., 2008; Hussong et al., 2007; Kaplow, Curran, Dodge, & The Conduct Problems Prevention Research Group, 2002). Each individual data set examined a relatively small age span, offering limited information specific to that sample. However, when the item-level data was pooled across the three studies, participant ages ranged from early childhood to adulthood, allowing researchers to examine a much broader trajectory of the development of substance use and disorder and more precisely understand how age moderates this development. They were also able to examine subsets of particular forms of parental alcoholism as moderators of youth substance use, not otherwise possible because of the low base-rates of these subgroups. Using IDA, they have examined a number of other new questions in these data sets, including how a number of other parent variables, not just alcoholism, predict a range of youth outcomes, including substance use and internalizing and externalizing problems. Pooling data has allowed them to examine original and new hypotheses across a wider range of variables with increased statistical power and more recent and appropriate statistical analyses.

However, these few studies appear to compose the majority of research in the behavioral sciences developing and applying IDA. One reason aggregating participant-level data is more common in fields such as medicine is the similarity in methodology across medical studies that make it easier to aggregate and analyze data (Cooper & Patall, 2009). The sparse use of IDA in the psychological literature is likely largely due to the difficulties associated with addressing the differences between heterogeneous samples. When combining participant-level data, differences between samples, measures and methodology cannot be ignored. In the alcoholism studies described above, the first necessary and significant step in aggregating the data was to identify which measures in each study were used to examine the constructs of interest (i.e., parental alcoholism and youth symptoms) and make the measures equivalent across studies. This step is arguably the most difficult and time-consuming step in IDA, but only after establishing that the constructs were measured the same way across the studies could they establish factor scores for each individual on the constructs and then use them in analyses across studies.

Aims of Current Study

The primary goal of the current study was to conduct this initial step of IDA using the 10 data sets focused on assessing the psychometric properties of the key measures used in the dissonance-based intervention for the prevention of eating disorders. My three research aims were as follows:

1. Gather and organize data from 10 heterogeneous data sets focused on dissonance-based interventions for eating disorders;
2. Identify the constructs of thin ideal internalization and body dissatisfaction in each study and identify how those constructs were measured;
3. Establish measurement invariance for each construct across studies.

This project constituted the first and necessary step in a larger project that will help further research regarding this dissonance-based intervention by allowing researchers to examine overall treatment effectiveness and moderators of treatment outcome across studies.

Method

Studies

I combined data sets from 10 studies focused on the dissonance-based intervention for the prevention of eating disorders (see Table 1). Stice and colleagues (2000, 2001, 2002, 2006, 2009; McMillan et al., 2011) conducted six studies, including one preliminary pilot study and five randomized trials, testing the effectiveness of a dissonance-based intervention in reducing thin ideal internalization and corresponding risk factors, such as body dissatisfaction. The six studies targeted female adolescents and young adults, ages 13-29 across studies, with significant weight concerns, who were at an increased risk for bulimic symptoms. Stice et al. generally used 3-week interventions and different comparison groups across the trials (i.e., dissonance vs.

assessment only control). Becker and colleagues conducted four studies (2005, 2006, 2008, & 2010) examining dissonance-based interventions. However, in these studies, the interventions was implemented over 2 weeks and focused on sorority members, who did not necessarily have weight concerns. Becker et al. also used different comparison groups across studies (i.e., dissonance vs. healthy weight management control). Most studies across both research groups used random assignment and also used a pre-test, post-test design with multiple follow-up points. For this study, only data from the first, pre-intervention time point in each study was used.

Table 1

Summary of Data Sets

Study	*n*	Age range	Thin ideal internalization	Body dissatisfaction	Treatment duration
Stice, Mazotti, Weibel, & Agras, 2000	30	18-22	IBSS-R	BDS BSQ	3 weeks
Stice, Chase, Stormer, & Appel, 2001	87	17-29	IBSS-R	BDS BSQ	3 weeks
Stice, Trost, & Chase, 2003	148	13-20	IBSS-R	BDS	3 weeks
Stice, Shaw, Burton, Wade, 2006	481	13-19	IBSS-R	BDS	3 weeks
Stice, Rohde, Gau, & Shaw, 2009	342	13-19	IBSS-R	BDS	4 weeks
McMillan, Stice, & Rohde, 2011	124	18-50	IBSS-R	BDS	4 weeks
Becker, Smith, Ciao, 2005	149	18-22	IBSS-R	BSQ	2 weeks
Becker, Smith, Ciao, 2006	90	18-20	IBSS-R	BSQ	2 weeks
Becker, Bull, Schaumberg, Cauble, Franco, 2008	188	18-21	IBSS-R	BSQ	2 weeks
Becker, Wilson, Williams, Kelly, McDaniel, & Elmquist, 2010	106	18-21	IBSS-R	**	2 weeks

Note: **This construct not available

Measures

Ideal Body Stereotype Scale-Revised (IBSS-R). The IBSS-R (Stice et al., 1996) was used in all 10 studies (see Table 1). The IBSS-R is an 8-item measure developed to assess thin ideal internalization. Participants responded to statements about the thin ideal (e.g., "Slim women are more attractive") on a 5-point Likert scale format with the anchors of "strongly disagree" to "strongly agree." Internal consistency estimates for the IBSS-R range between .89 and .91 and 2-week test-retest of .80 (Stice & Agras, 1998; Stice, Ziemba, et al., 1996; Stice, Fisher, & Martinez, 2004).

Satisfaction and Dissatisfaction with Body Part Scale (BDS). The nine items on the BDS were taken from the longer Satisfaction and Dissatisfaction with Body Parts Scale (Berscheid, Walster, & Bohrnstedt, 1973) and were used in all six of Stice's studies to assess body dissatisfaction. On the BDS, respondents indicated how satisfied they were with nine different aspects of their body (such as "weight" or "legs"). Internal consistency has been estimated at .94 and 3-week test-retest at .90 (Stice & Agras, 1998; Stice, Ziemba et al., 1996). Although the BDS was used in five studies, there were some differences in the item stems ("Over the past week" vs. "Over the past month") across studies.

Body Shape Questionnaire (BSQ). A 15-item version of the BSQ (Cooper, Taylor, Cooper, & Fairburn, 1987) was used in two of Stice's studies and a 34-item version was used in three of Becker's studies used to assess body dissatisfaction. Only the 15 BSQ items common to all studies were used for the purposes of this study. Participants responded to questions on a 6-point Likert scale with anchors of "never" and "always" (sample item "How many times over the last week have you felt ashamed of your body?"). Past research has estimated that the internal consistency for the BSQ is .93 and 3-week test-retest reliability is .88 (Dowson & Henderson, 2001; Rosen, Jones, Ramirez, & Waxman, 1996).

Analysis

Once the data were organized and combined into one data set, I fit measurement invariance models for all three measures, based on individual item responses from the pre-intervention time point of each study. Establishing measurement invariance refers to evaluating a measure and ensuring that the measurement properties remain the same across groups of interest, such as gender, geographic region, or ethnicity (Brown, 2006). If the measurement properties are not the same across groups, or if the measure is not assessing the construct the same way across groups, then comparing scale or factor scores across groups is challenging because they may mean different things for different groups. If the measurement properties are the same across groups, measurement invariance has been established and scores can be compared across groups. Ideally all items would be completely invariant across all groups, but in this study there were some items that were not fully invariant, resulting in partial measurement variance (Byrne, Shavelson, & Muthén, 1989). However, partial measurement invariance still allows us to examine how items differ across groups and compare scores across groups. For the purposes of this study, I also examined measurement invariance across different researchers. All confirmatory factor analysis models were analyzed using Mplus Version 7.

Thin ideal internalization. I followed a standard multiple groups confirmatory factor analysis procedure as the first part of establishing measurement invariance in the IBSS-R (Brown, 2006). First, I fit a one-factor solution within each study, with all eight items loading on a thin ideal internalization factor. I then tested a common factor structure (configural invariance), establishing that the one factor solution provided a good fit across studies. To identify the configural model, the factor variance and mean were set to 1 and 0, respectively, for the first study, and the loading and intercept for the first item were constrained to be equal across studies. The latent factor means were freed for all but the first study. After establishing configural invariance, I systematically tested the equality of different parameters, successively constraining additional parameters to be equal. I began by constraining the factor loadings (metric invariance) to be equal, testing the equality of the relationships between the indicators and latent factor across studies. I then proceeded to constrain the intercepts (indicator means) to examine scalar invariance. In cases where the factor loadings were freed in the metric models, those loading parameters and corresponding intercepts were also freed in the subsequent scalar invariance models. This process was part of assessing differential item functioning (DIF), or whether individuals who had equal values on the latent constructs responded similarly to each item across studies. I used likelihood-ratio tests (LRT) with degrees of freedom equal to the difference in the number of parameters in the models to compare the fit of competing models.

Body dissatisfaction. Establishing measurement invariance across the body dissatisfaction construct was more challenging because of the two different measures (BDS and BSQ) used to assess it. Previous studies combining different measures have typically used relatively straightforward, often dichotomous, data that could be harmonized relatively easily. One contribution of this project was to explore ways of combining somewhat more challenging methodologies across a larger number of studies. I used the following procedure to explore and address this. First, I tried using a process called "chaining" to establish invariance across body dissatisfaction and place the two measures in a common metric across studies (Bauer & Hussong, 2009). Chaining is a way to connect studies that use different measures through items common to multiple studies. Because two studies used both the BSQ and BDS, I originally hoped that they would serve as a link between studies that used only the BDS and others that used only the BSQ. Chaining was the optimal option because it would not require any adaptation of the measures. However, I was not able to use chaining as it is currently described in the literature for a number of reasons. One primary challenge I faced was that the two studies using both measures had small sample sizes and one (Stice 2000) ultimately had to be dropped from analyses. I also encountered challenges due to limitations in available statistical software in fitting models across so many studies with so few individuals with data on both measures. I did not attempt harmonization for this data due to the high degree of interrelatedness of items within measures, which made it difficult to distinguish why one item was better to use than another for harmonization purposes. Harmonization also proved infeasible given the two distinct approaches between measures to assess body dissatisfaction. Due to the difficulty in chaining the measures, I analyzed the measurement invariance for each measure separately, using all the studies that used those constructs. First, I established a one-factor model for each study within each study. I then conducted factor analyses across studies, holding parameters equal across studies and systematically testing model parameters across studies.

Age effects. I originally proposed examining age as a moderator of factor loadings and factor scores. However, age was largely confounded with the study variables, as seen in Table 1. Thus, after establishing measurement invariance across study, I examined whether age predicted the latent values for thin ideal internalization and body dissatisfaction.

Results

Thin Ideal Internalization

Preliminary analyses. Preliminary analyses of the eight items on the IBSS-R suggested items were distributed relatively normally, although most items showed mild negative skew.

Initial factor analysis of eight items showed that items 7 ("Curvy women are more attractive") and 8 ("Shapely women are more attractive") loaded significantly more poorly than other items across all studies. Reliability across the pooled studies increased from .81 to .84 when the two items were dropped from the analyses. Within studies, reliability also increased for every study when items 7 and 8 were dropped, particularly within the Stice studies. These results suggest that "curvy" and "shapely" may not characterize the thin ideal as well as the other items, particularly for younger populations. Based on these analyses, items 7 and 8 were dropped from subsequent analyses.

Measurement invariance. I followed a standard multiple groups factor analysis procedure to establish measurement invariance of the IBSS-R (Brown, 2006). First, I fit a one-factor solution within each study, with the six remaining items loading onto the thin ideal internalization factor. Modification indices indicated correlations between items 1 ("Slim women are more attractive") and 5 ("Slender women are more attractive"), items 2 ("Tall women are more attractive") and 6 ("Women with long legs are more attractive"), and items 3 ("Women with toned bodies are more attractive") and 4 ("Women in shape are more attractive"). These three correlated errors were added to the model based on modification indices and the semantic similarities between items; the addition of these correlated errors improved the fit of the model across all studies (see Figure 1).

Factor loadings and goodness of fit for each individual study can be found in Figure 1 and Table 2. Of note, with these additional parameters added to the model, the sample from Stice 2000 was no longer large enough to allow for an identified model (N = 30). This study was dropped from further analyses.

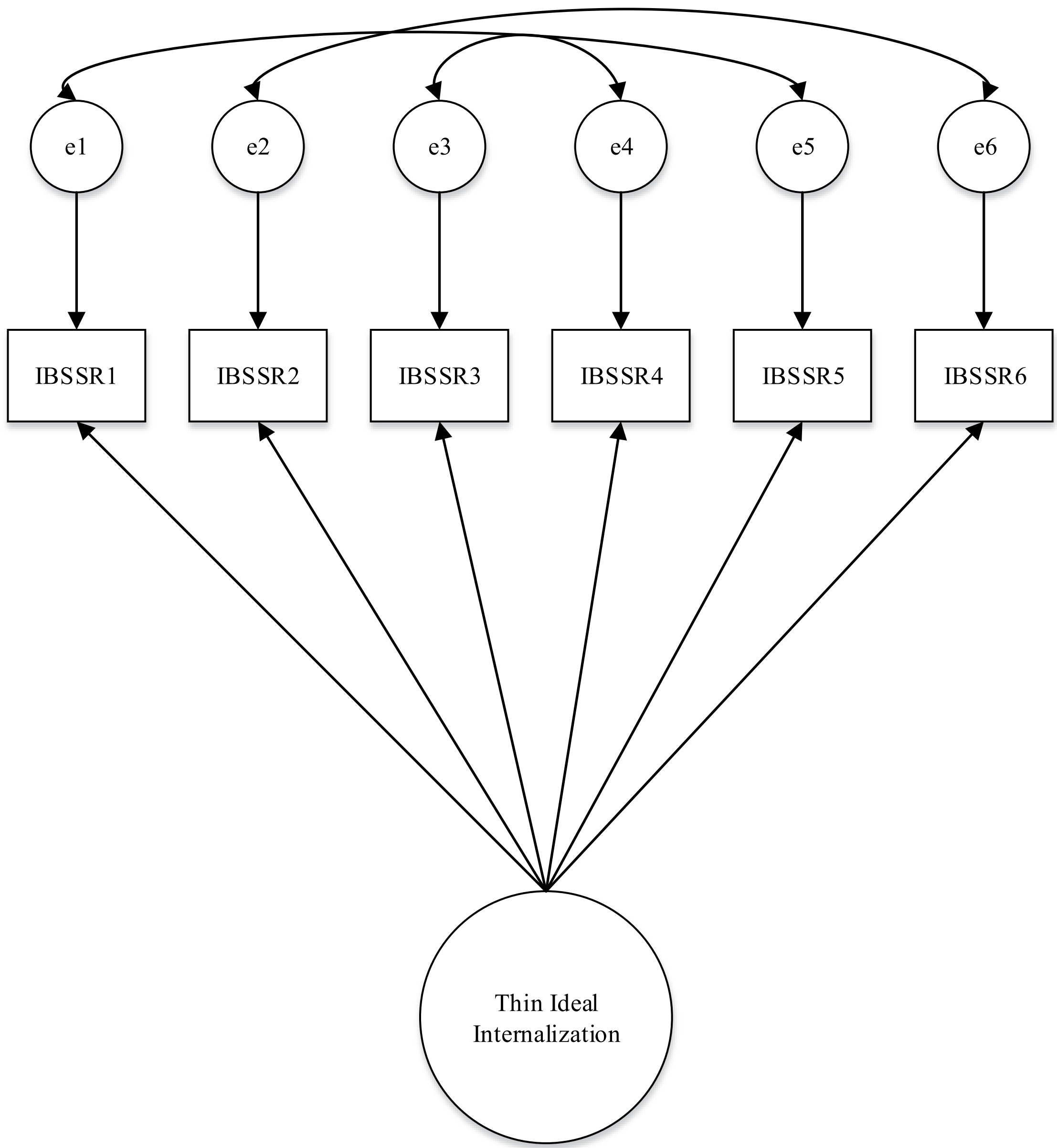

Figure 1. One-factor solution for the IBSS-R tested within and across studies.

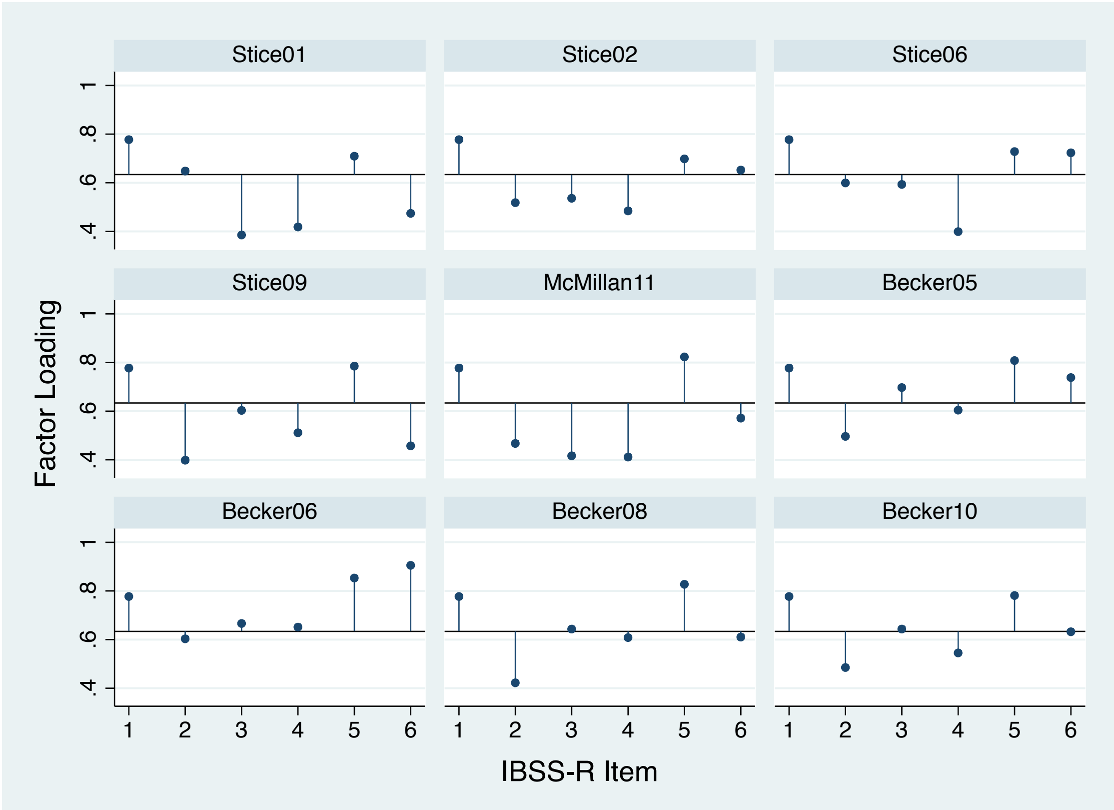

Figure 2. Factor loadings of the IBSS-R shown by study around the average factor loading across items.

Table 2

Factor Loadings and Fit for One-Factor Solution Within Studies for IBSS-R

	Study								
	Stice 2001	Stice 2002	Stice 2006	Stice 2009	McMillan 2011	Becker 2005	Becker 2006	Becker 2008	Becker 2010
IBSS-R Item									
Item 1	0.78	0.78	0.64	0.68	0.71	0.71	0.68	0.70	0.76
Item 2	0.65	0.42	0.49	0.35	0.43	0.45	0.53	0.38	0.48
Item 3	0.38	0.42	0.49	0.53	0.38	0.63	0.58	0.58	0.63
Item 4	0.42	0.38	0.33	0.45	0.38	0.55	0.57	0.54	0.53
Item 5	0.71	0.70	0.60	0.68	0.76	0.73	0.75	0.74	0.77
Item 6	0.47	0.50	0.59	0.40	0.53	0.67	0.79	0.55	0.62
Model Fit									
RMSEA	0.10	0.08	0.04	0.07	0.08	0.09	0.00	0.00	0.00
CFI	0.98	0.98	1.00	0.99	0.99	0.99	1.00	1.00	1.00
TLI	0.95	0.96	0.99	0.97	0.97	0.97	1.03	1.01	1.03

After fitting a one-factor solution within each study, configural invariance of the solution was tested across models. The covariance between items 1 and 5 was constrained to 0 in one study where the point estimate was close to 0 but negative, impairing model estimation. This model, with the rest of factor loadings and means estimated freely, showed good model fit (see Tables 3-5). After establishing configural invariance, all factor loadings were constrained to be equal across groups. The metric invariance model also showed good fit.

Table 3

Factor Loadings for the IBSS-R Across Models

	Model				
	Configural	Metric	Partial Metric	Scalar	Partial Scalar
Stice01					
Item 1	0.78	0.68	0.64	0.63	0.64
Item 2	0.65	0.44	0.67*	0.66*	0.66*
Item 3	0.39	0.52	0.49	0.49	0.49
Item 4	0.42	0.45	0.45	0.45	0.45
Item 5	0.71	0.68	0.64	0.65	0.65
Item 6	0.47	0.57	0.55	0.55	0.55
Stice02					
Item 1	0.78	0.68	0.64	0.63	0.64
Item 2	0.52	0.44	0.41	0.40	0.40
Item 3	0.54	0.52	0.49	0.49	0.49
Item 4	0.48	0.45	0.45	0.45	0.45
Item 5	0.70	0.68	0.64	0.65	0.65
Item 6	0.65	0.57	0.55	0.55	0.55
Stice06					
Item 1	0.78	0.68	0.64	0.63	0.64
Item 2	0.60	0.44	0.41	0.40	0.40
Item 3	0.59	0.52	0.49	0.49	0.49
Item 4	0.40	0.45	0.33*	0.33*	0.33*
Item 5	0.73	0.68	0.64	0.65	0.65
Item 6	0.72	0.57	0.55	0.55	0.55

(continued)

Stice09					
Item 1	0.78	0.68	0.64	0.63	0.64
Item 2	0.40	0.44	0.41	0.40	0.40
Item 3	0.60	0.52	0.49	0.49	0.49
Item 4	0.51	0.45	0.45	0.45	0.45
Item 5	0.79	0.68	0.64	0.65	0.65
Item 6	0.46	0.57	0.43*	0.43*	0.43*
McMillan11					
Item 1	0.78	0.68	0.64	0.63	0.64
Item 2	0.47	0.44	0.41	0.40	0.40
Item 3	0.42	0.52	0.49	0.49	0.49
Item 4	0.41	0.45	0.45	0.45	0.45
Item 5	0.82	0.68	0.64	0.65	0.65
Item 6	0.57	0.57	0.55	0.55	0.55
Becker05					
Item 1	0.78	0.68	0.64	0.63	0.64
Item 2	0.50	0.44	0.41	0.40	0.40
Item 3	0.70	0.52	0.49	0.49	0.49
Item 4	0.60	0.45	0.45	0.45	0.45
Item 5	0.81	0.68	0.64	0.65	0.65
Item 6	0.74	0.57	0.55	0.55	0.55
Becker06					
Item 1	0.78	0.68	0.64	0.63	0.64
Item 2	0.60	0.44	0.41	0.40	0.40
Item 3	0.67	0.52	0.49	0.49	0.49
Item 4	0.65	0.45	0.45	0.45	0.45
Item 5	0.85	0.68	0.64	0.65	0.65
Item 6	0.91	0.57	0.55	0.55	0.55
Becker08					
Item 1	0.78	0.68	0.64	0.63	0.64
Item 2	0.42	0.44	0.41	0.40	0.40
Item 3	0.64	0.52	0.49	0.49	0.49
Item 4	0.61	0.45	0.45	0.45	0.45
Item 5	0.83	0.68	0.64	0.65	0.65
Item 6	0.61	0.57	0.55	0.55	0.55
Becker10					
Item 1	0.78	0.68	0.64	0.63	0.64
Item 2	0.49	0.44	0.41	0.40	0.40
Item 3	0.64	0.52	0.49	0.49	0.49
Item 4	0.55	0.45	0.45	0.45	0.45
Item 5	0.78	0.68	0.64	0.65	0.65
Item 6	0.63	0.57	0.55	0.55	0.55

Note. * denotes freely estimated parameters in partial metric and later models

Table 4

Item Intercepts for the IBSS-R Across Models

	Model				
	Configural	Metric	Partial Metric	Scalar	Partial Scalar
Stice01					
Item 1	3.75	3.75	3.75	3.90	3.76**
Item 2	3.38	3.38	3.38	3.35*	3.35*
Item 3	4.18	4.18	4.18	4.22	4.25
Item 4	4.35	4.35	4.35	4.33	4.35
Item 5	3.89	3.89	3.89	3.90	3.91
Item 6	3.66	3.66	3.66	3.55	3.56
Stice02					
Item 1	3.75	3.75	3.75	3.90	3.81**
Item 2	3.26	3.26	3.26	3.17	3.31**
Item 3	4.14	4.14	4.14	4.22	4.25
Item 4	4.30	4.30	4.31	4.33	4.35
Item 5	3.86	3.86	3.86	3.90	3.91
Item 6	3.49	3.49	3.49	3.55	3.56
Stice06					
Item 1	3.75	3.75	3.75	3.90	3.92
Item 2	3.05	3.07	3.07	3.17	3.17
Item 3	4.04	4.04	4.04	4.22	4.17**
Item 4	4.28	4.26	4.28	4.39*	4.37*
Item 5	3.77	3.76	3.76	3.90	3.91
Item 6	3.36	3.38	3.37	3.55	3.56
Stice09					
Item 1	3.75	3.75	3.75	3.90	3.92
Item 2	3.05	3.09	3.08	3.17	3.17
Item 3	4.00	3.99	3.99	4.22	4.14**
Item 4	4.23	4.23	4.24	4.33	4.37**
Item 5	3.70	3.70	3.70	3.90	3.91
Item 6	3.36	3.42	3.38	3.49*	3.49*
McMillan11					
Item 1	3.75	3.75	3.75	3.90	3.92
Item 2	3.03	3.02	3.03	3.17	3.17
Item 3	4.24	4.23	4.23	4.22	4.25
Item 4	4.34	4.34	4.34	4.33	4.35
Item 5	3.76	3.76	3.76	3.90	3.91
Item 6	3.50	3.50	3.50	3.55	3.56

(continued)

Becker05					
Item 1	3.75	3.75	3.75	3.90	3.92
Item 2	3.01	3.01	3.01	3.17	3.17
Item 3	4.20	4.21	4.21	4.22	4.35**
Item 4	4.19	4.20	4.20	4.33	4.33**
Item 5	3.72	3.72	3.72	3.90	3.91
Item 6	3.41	3.43	3.42	3.55	3.56
Becker06					
Item 1	3.75	3.75	3.75	3.90	3.92
Item 2	2.98	2.97	2.97	3.17	3.17
Item 3	4.05	4.05	4.05	4.22	4.25
Item 4	4.17	4.16	4.16	4.33	4.35
Item 5	3.68	3.67	3.67	3.90	3.91
Item 6	3.47	3.45	3.45	3.55	3.56
Becker08					
Item 1	3.75	3.75	3.75	3.90	3.92
Item 2	3.07	3.08	3.08	3.17	3.17
Item 3	4.13	4.12	4.12	4.22	4.25
Item 4	4.19	4.18	4.19	4.33	4.35
Item 5	3.70	3.69	3.69	3.90	3.91
Item 6	3.34	3.35	3.35	3.55	3.56
Becker10					
Item 1	3.75	3.75	3.75	3.90	3.92
Item 2	2.94	2.95	2.95	3.17	3.17
Item 3	4.15	4.14	4.14	4.22	4.25
Item 4	4.18	4.18	4.18	4.33	4.35
Item 5	3.73	3.73	3.73	3.90	3.91
Item 6	3.42	3.42	3.43	3.55	3.56

Note. * denotes parameters freed from metric model; ** denotes parameters freed in partial scalar model

Table 5

Model Fit for the IBSS-R Across Studies

Model	RMSEA	CFI	TLI	χ^2	df
Configural	0.05	0.99	0.99	82.85	55.00
Metric	0.05	0.99	0.99	133.68	95.00
Partial Metric	0.03	1.00	0.99	109.72	92.00
Scalar	0.05	0.99	0.99	176.52	129.00
Partial Scalar	0.03	1.00	1.00	138.09	121.00

The Likelihood-Ratio Test (LRT) between the two models was not significant, indicating that constraining factor loadings did not result in a significant decrease in model fit ($\chi^2(40) = 50.83$, $p = .12$; Bauer, 2009). However, modification indices suggested relatively large expected parameter change and improved fit with the loadings for item 2 ("Tall women are more attractive") in Stice 2001, item 4 ("Women who are in shape are more attractive") in Stice 2006, and item 6 ("Women with long legs are more attractive") in Stice 2009 (see Table 3 and Figure 1). A revised metric model showed excellent fit and the LRT test remained nonsignificant, $\chi^2(37) = 26.87$, $p = .89$, indicating partial metric invariance. A scalar invariance model, where item loadings and means were constrained to be equal across all groups (except freed parameters from the partial metric model), also showed good model fit. However, the LRT indicated a significant difference between the partial metric and scalar models, $\chi^2(37) = 66.80$, $p < .01$).

Modification indices were used to determine areas of misfit and parameters with the greatest expected change were systematically released and the models were compared again. Based on this process, eight different item means were freed in various studies (see Table 4). Although the means were only minimally higher in these studies compared to others (see Figure 3), an LRT test was no longer significant after freely estimating these parameters, $\chi^2(29) = 28.37$, $p = .50$.

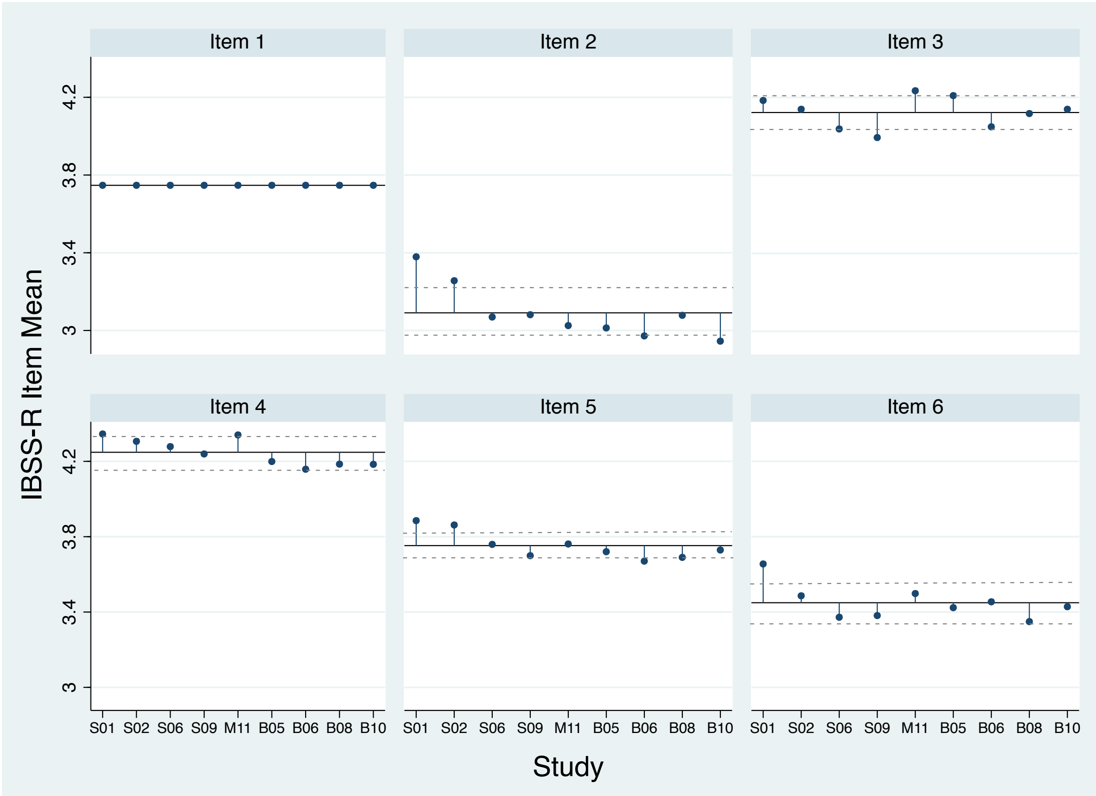

Figure 3. IBSS-R item means shown around the item mean across studies for each respective item. Dashed lines represent one standard deviation above and below each item mean.

Body Dissatisfaction

Preliminary analyses. Preliminary analyses of the BDS suggested that items were distributed relatively normally, with slight positive skew on the first eight and slight negative skew on item 9. Initial factor analysis showed that item 9 ("Legs") loaded significantly more poorly than other items across all studies. Reliability across the combined data set increased overall from .87 to .90 when this item was dropped from the analysis. These results suggest that individuals perceive a different relationship between overall body dissatisfaction and "legs" than the other eight items. Based on these analyses, item 9 was dropped from subsequent analyses.

For the purposes of this study, the 15 BSQ items common to all studies were examined for measurement invariance. Reliability for the 15 items was .93. Preliminary analyses indicated that most items were relatively normally distributed, although five items showed moderate positive skew. As in the analyses of the IBSS-R, the sample from Stice 2000 was no longer large enough to allow for an identified model in either the BDS or BSQ and was dropped from the analyses of both body dissatisfaction measures.

Measurement invariance. As with the IBSS-R, I followed a standard multiple groups factor analysis procedure to establish measurement invariance of the BDS and BSQ. As noted above, this process was followed for each of the measures separately.

BDS. Initial factor analyses for the BDS within and across studies suggested residual correlations between six items with similar semantic characteristics. These correlated errors were added to the model and resulted in improved model fit across all studies. Factor loadings and goodness of fit for each individual study can be found in Table 6. The one-factor solution with the correlated errors can be seen in Figure 4.

Table 6

Factor Loadings and Fit for One-Factor Solution Within Studies for BDS

	Factor Loadings				
	Stice 2001	Stice 2002	Stice 2006	Stice 2009	McMillan 2011
BDS Item					
Item 1	1.12	1.09	0.91	0.86	1.01
Item 2	0.98	0.90	0.88	0.85	0.85
Item 3	0.69	0.72	0.78	0.80	0.68
Item 4	0.64	0.76	0.76	0.69	0.76
Item 5	0.83	0.91	0.81	0.81	0.83
Item 6	0.66	0.77	0.57	0.77	0.60
Item 7	0.66	0.72	0.67	0.61	0.46
Item 8	0.75	0.74	0.72	0.71	0.70
Model Fit					
RMSEA	0.07	0.09	0.07	0.08	0.09
CFI	0.98	0.98	0.98	0.98	0.97
TLI	0.97	0.95	0.97	0.96	0.94

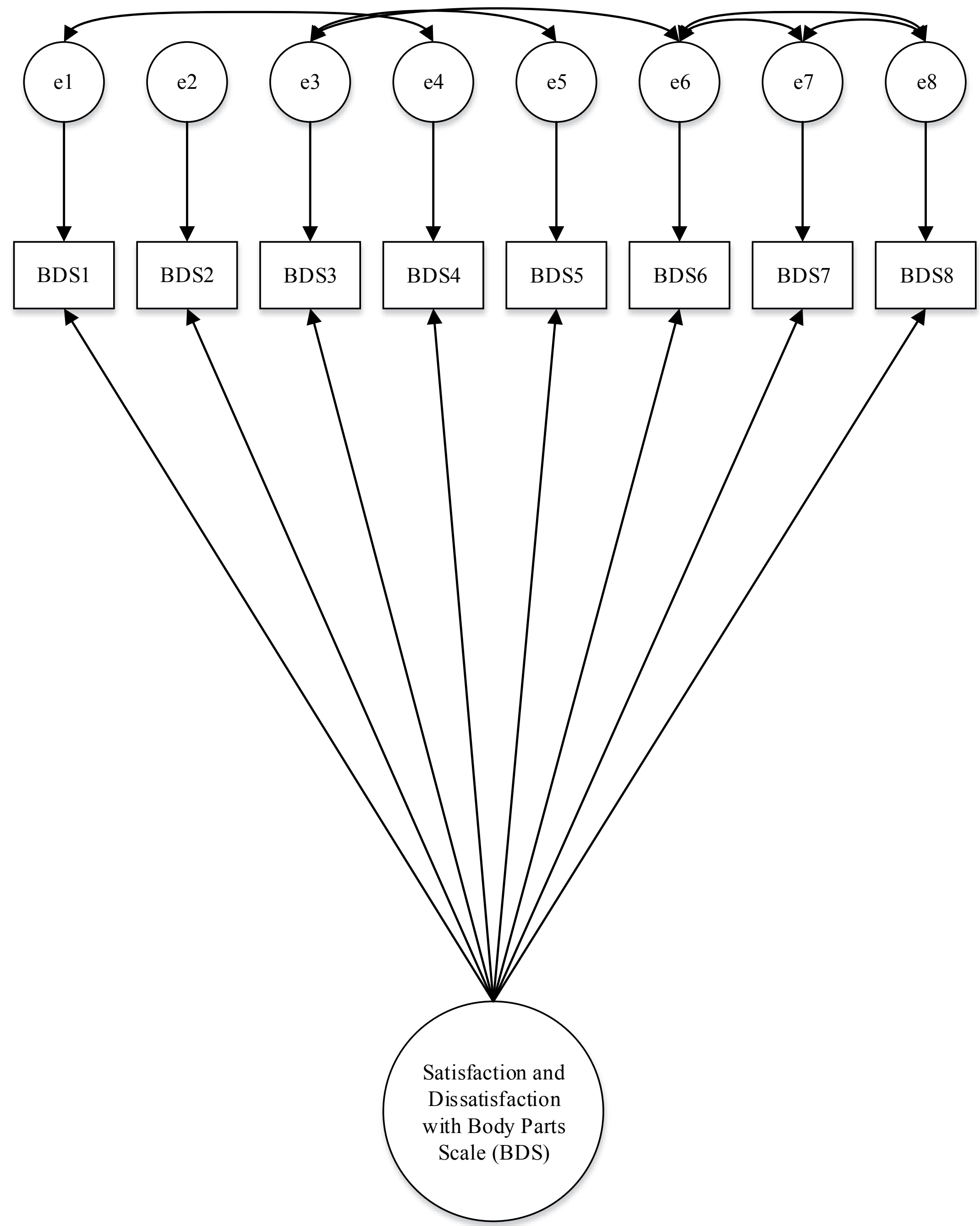

Figure 4. One-factor solution for the BDS within and across studies.

After establishing configural invariance across studies, all factor loadings were constrained to be equal across groups. Factor loadings and model fit across studies can be seen in Figure 5 and Table 7. Factor loadings and intercepts across models can be found in Tables 8 and 9, respectively. The metric model showed relatively good fit and the LRT test between the metric and configural models was nonsignificant, $\chi^2(28) = 40.07$, $p = .07$. However, modification indices indicated large expected parameter changes if the item loading for item 6 ("Thighs") was released in Stice 2006 and 2009. With these parameters released, the model showed better fit and the LRT test remained nonsignificant, $\chi^2(26) = 27.43$, $p = .39$, indicating partial metric invariance. A scalar invariance model, where item loadings and means were constrained to be equal across all groups, also showed good model fit. However, the LRT test indicated significantly degraded fit between the partial metric and scalar models, $\chi^2(26) = 429.05$, $p < .001$. As in previous analyses, because item 6 in Stice 2006 and 2009 were freely estimated in the metric model, that loading and intercept were also freed in the scalar model.

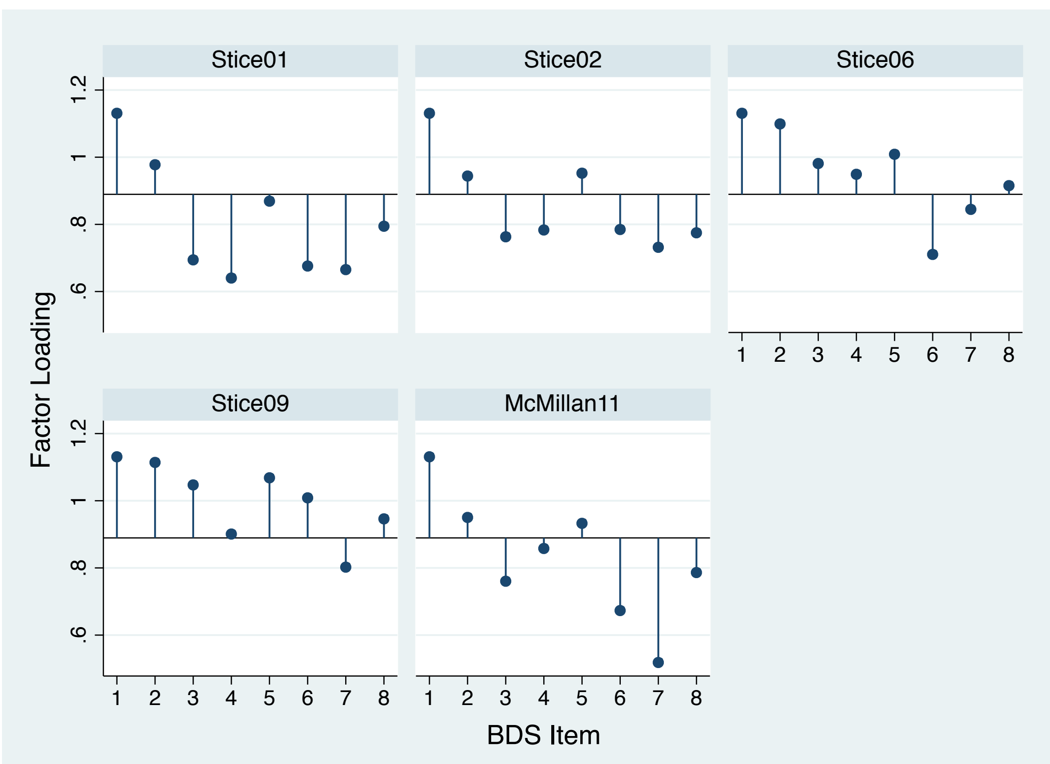

Figure 5. Factor loadings of the BDS shown by study around the average factor loading across items.

Table 7

Model Fit for the BDS

Model	RMSEA	CFI	TLI	χ^2	df
Configural	0.08	0.98	0.96	164.20	70.00
Metric	0.07	0.98	0.97	204.27	98.00
Partial Metric	0.07	0.98	0.97	191.63	96.00
Scalar	0.14	0.89	0.88	620.68	122.00
Partial Scalar	0.06	0.98	0.97	209.81	111.00

Table 8

Factor Loadings for the BDS Across Models

	Model				
	Configural	Metric	Partial Metric	Scalar	Partial Scalar
Stice01					
Item 1	1.13	1.04	1.04	1.02	1.05
Item 2	0.98	0.96	0.96	0.93	0.95
Item 3	0.69	0.83	0.83	1.01	0.84
Item 4	0.64	0.80	0.80	0.75	0.78
Item 5	0.87	0.91	0.91	0.91	0.93
Item 6	0.68	0.73	0.73	0.82	0.72
Item 7	0.67	0.70	0.70	0.70	0.69
Item 8	0.79	0.80	0.80	0.79	0.80
Stice02					
Item 1	1.13	1.04	1.04	1.02	1.05
Item 2	0.94	0.96	0.96	0.93	0.95
Item 3	0.76	0.83	0.83	1.01	0.84
Item 4	0.78	0.80	0.80	0.75	0.78
Item 5	0.95	0.91	0.91	0.91	0.93
Item 6	0.78	0.73	0.73	0.82	0.72
Item 7	0.73	0.70	0.70	0.70	0.69
Item 8	0.77	0.80	0.80	0.79	0.80
Stice06					
Item 1	1.13	1.04	1.04	1.02	1.05
Item 2	1.10	0.96	0.96	0.93	0.95
Item 3	0.98	0.83	0.83	1.01	0.84
Item 4	0.95	0.80	0.80	0.75	0.78
Item 5	1.01	0.91	0.91	0.91	0.93
Item 6	0.71	0.73	0.62*	0.61*	0.61*
Item 7	0.84	0.70	0.70	0.70	0.69
Item 8	0.92	0.80	0.80	0.79	0.80

(continued)

Stice09					
Item 1	1.13	1.04	1.04	1.02	1.05
Item 2	1.11	0.96	0.96	0.93	0.95
Item 3	1.05	0.83	0.83	1.01	0.84
Item 4	0.90	0.80	0.80	0.75	0.78
Item 5	1.07	0.91	0.91	0.91	0.93
Item 6	1.01	0.73	0.88*	0.86*	0.88*
Item 7	0.80	0.70	0.70	0.70	0.69
Item 8	0.95	0.80	0.80	0.79	0.80
McMillan11					
Item 1	1.13	1.04	1.04	1.02	1.05
Item 2	0.95	0.96	0.96	0.93	0.95
Item 3	0.76	0.83	0.83	1.01	0.84
Item 4	0.86	0.80	0.80	0.75	0.78
Item 5	0.93	0.91	0.91	0.91	0.93
Item 6	0.67	0.73	0.73	0.82	0.72
Item 7	0.52	0.70	0.70*	0.70*	0.69*
Item 8	0.79	0.80	0.80	0.79	0.80

Note. * denotes freely estimated parameters in partial metric and later models

Table 9

Item Intercepts for the BDS Across Models

	Model				
	Configural	Metric	Partial Metric	Scalar	Partial Scalar
Stice01					
Item 1	3.31	3.31	3.31	3.39	3.33
Item 2	3.30	3.30	3.30	3.33	3.27
Item 3	3.84	3.84	3.84	3.42	3.81**
Item 4	3.19	3.19	3.19	3.27	3.23
Item 5	3.31	3.31	3.31	3.33	3.25
Item 6	3.76	3.76	3.76	3.54	3.76
Item 7	3.69	3.69	3.69	3.23	3.69**
Item 8	3.41	3.41	3.41	3.22	3.41**
Stice02					
Item 1	3.31	3.31	3.31	3.39	3.33
Item 2	3.24	3.25	3.25	3.33	3.27
Item 3	3.63	3.65	3.65	3.42	3.68**
Item 4	3.16	3.17	3.17	3.27	3.23
Item 5	3.22	3.22	3.22	3.33	3.25
Item 6	3.82	3.82	3.82	3.54	3.84**
Item 7	3.51	3.51	3.51	3.23	3.53**
Item 8	3.29	3.31	3.31	3.22	3.33**
Stice06					
Item 1	3.31	3.31	3.31	3.39	3.33
Item 2	3.31	3.28	3.28	3.33	3.27
Item 3	2.84	2.79	2.79	3.42	2.80**
Item 4	3.33	3.27	3.27	3.27	3.23
Item 5	3.24	3.23	3.23	3.33	3.25
Item 6	2.64	2.70	2.62	2.70*	2.61*
Item 7	3.16	3.10	3.10	3.23	3.07
Item 8	3.17	3.14	3.14	3.22	3.13
Stice09					
Item 1	3.31	3.31	3.31	3.39	3.33
Item 2	3.18	3.18	3.18	3.33	3.27
Item 3	3.75	3.76	3.76	3.42	3.81**
Item 4	3.12	3.12	3.12	3.27	3.23
Item 5	3.24	3.24	3.24	3.33	3.25
Item 6	3.51	3.52	3.51	3.68*	3.58*
Item 7	2.99	2.99	2.99	3.23	3.07
Item 8	3.04	3.05	3.05	3.22	3.13
McMillan11					
Item 1	3.31	3.31	3.31	3.39	3.33
Item 2	3.35	3.41	3.41	3.33	3.27

(continued)

Item 3	2.57	2.67	2.67	3.42	2.62**
Item 4	3.41	3.43	3.42	3.27	3.23
Item 5	3.19	3.23	3.23	3.33	3.25
Item 6	2.84	2.93	2.92	3.54	2.83**
Item 7	3.19	3.36	3.36	3.23	3.25**
Item 8	3.21	3.27	3.27	3.22	3.13

Note. * denotes parameters freed from metric model; ** denotes parameters freed in partial scalar model

Modification indices for the scalar model indicated misfit across studies on a number of the item intercepts, suggesting differences between mean values between studies; of note, the raw item mean differences were fairly small, as seen in Figure 6. A revised model releasing the mean for items 3 ("Appearance of the stomach") and 6 ("Thighs") across studies, as well as the mean for items 7 ("Buttocks) and 8 ("Legs") in several studies (see Table 9), showed good model fit and the LRT test was nonsignificant, $\chi^2(15) = 18.19$, $p = .25$. Overall, these analyses indicated partial measurement invariance of the BDS.

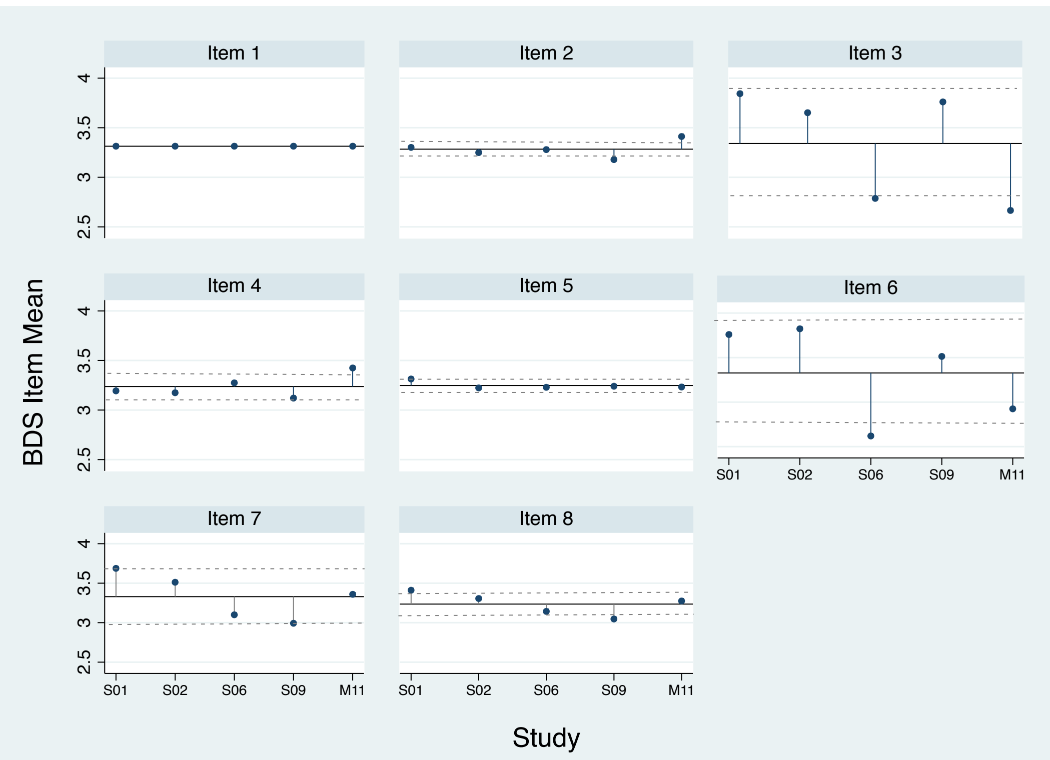

Figure 6. BDS item means by study shown around the item mean across studies for each respective item. Dashed lines represent one standard deviation above and below each item mean.

BSQ. Initial factor analyses within and across studies indicated correlations between four items with semantic similarities (see Figure 7). These correlated errors were added to the model and resulted in good model fit across all studies (see Table 10).

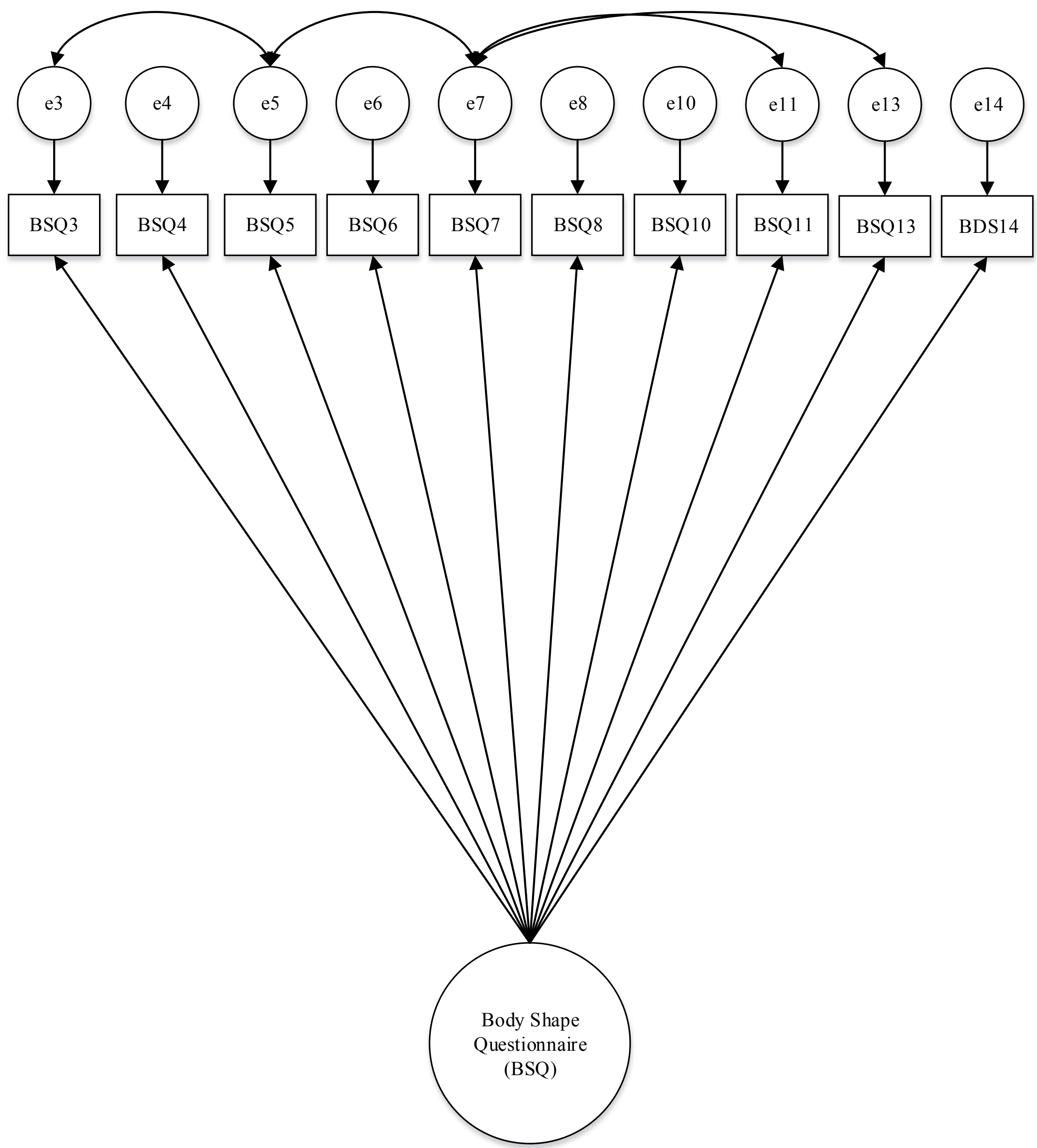

Figure 7. One-factor solution for the BSQ within and across studies.

Table 10

Factor Loadings and Fit for One-Factor Solution Within Studies for the BSQ

	Study			
	Stice 2001	Becker 2005	Becker 2006	Becker 2008
BSQ Item				
Item 3	0.88	0.72	0.68	0.69
Item 4	1.23	1.13	1.16	1.05
Item 5	1.08	0.59	0.33	0.63
Item 6	1.17	0.89	1.12	0.91
Item 7	1.02	0.44	0.38	0.55
Item 8	1.24	1.20	1.03	1.13
Item 10	0.86	0.59	0.50	0.75
Item 11	1.25	1.07	0.99	1.14
Item 13	1.24	0.96	0.92	0.98
Item 14	1.21	1.05	1.03	1.15
Model Fit				
RMSEA	0.07	0.04	0.03	0.08
CFI	0.98	0.99	0.99	0.97
TLI	0.97	0.99	0.99	0.96

After establishing configural invariance, all factor loadings were constrained to be equal across groups. Model fit and factor loadings across studies can be seen in Table 11 and Figure 8, respectively. Factor loadings and intercepts across models can be found in Tables 12 and 13, respectively. The metric model showed good fit but the LRT test between the metric and configural models indicated the additional constraints significantly degraded model fit, $\chi^2(27) = 51.63$, $p < .001$. Based on modification indices, the loadings for Item 7 ("Have you not gone out to social events/occasions (e.g., parties) because you have felt bad about your shape?") in Stice 2001 and items 5 ("Has thinking about your shape interfered with your ability to concentrate (e.g., while watching television)?") and 6 ("Have you avoided wearing clothes that make you particularly aware of the shape of your body?") in Becker 2006 were released. The revised model maintained good fit and the LRT test was no longer significant, $\chi^2(24) = 29.14$, $p = .21$, indicating partial metric invariance.

Table 11

Model Fit for the BSQ

Model	RMSEA	CFI	TLI	χ^2	df
Configural	0.06	0.98	0.97	178.753	124
Metric	0.07	0.97	0.97	230.387	151
Partial Metric	0.06	0.98	0.98	207.897	148
Scalar	0.07	0.97	0.96	275.398	172
Partial Scalar	0.06	0.98	0.98	231.155	165

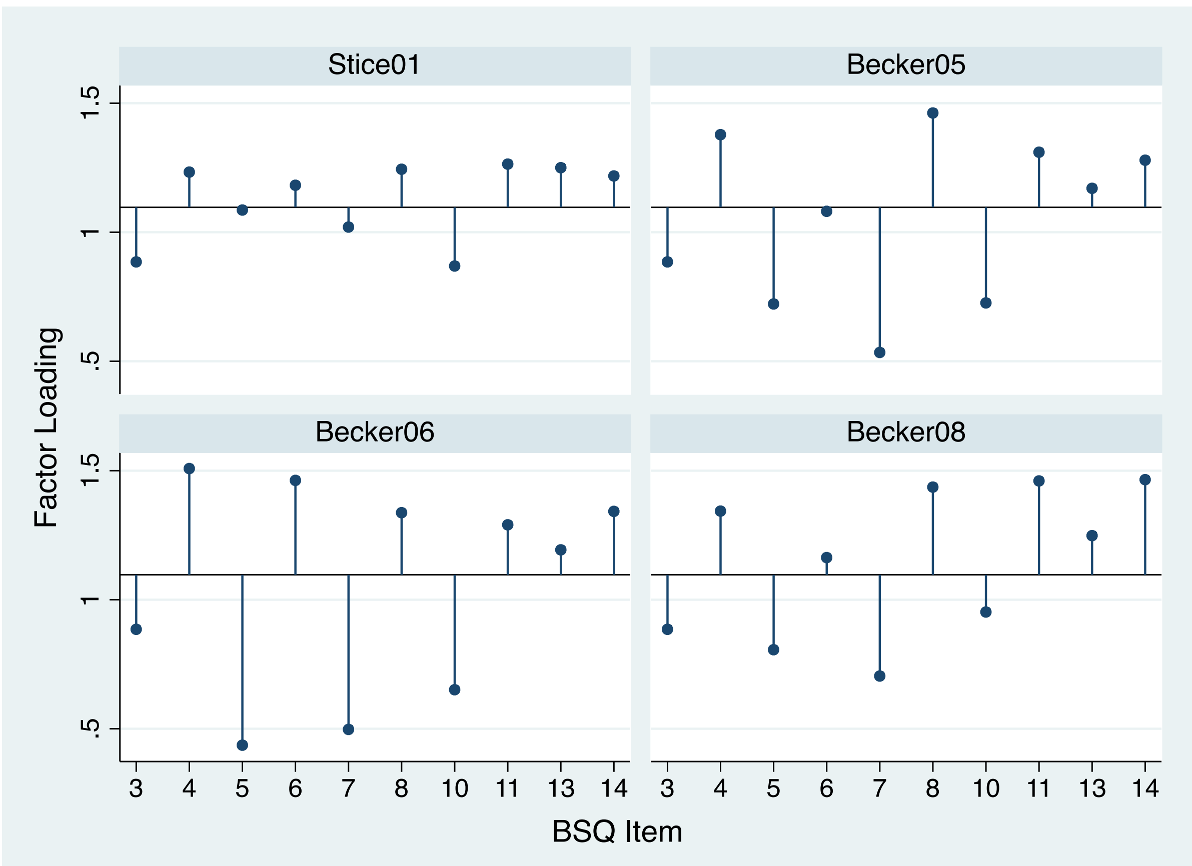

Figure 8. Factor loadings of the BSQ shown by study around the average factor loading across items.

Table 12

Factor Loadings for the BSQ Across Models

	Model				
	Configural	Metric	Partial Metric	Scalar	Partial Scalar
Stice01					
Item 3	0.89	0.86	0.86	0.85	0.86
Item 4	1.23	1.30	1.29	1.25	1.27
Item 5	1.09	0.71	0.77	0.79	0.78
Item 6	1.18	1.15	1.11	1.12	1.12
Item 7	1.02	0.60	0.90*	0.90*	0.90*
Item 8	1.24	1.34	1.33	1.32	1.32
Item 10	0.87	0.79	0.78	0.78	0.78
Item 11	1.26	1.33	1.29	1.30	1.30
Item 13	1.25	1.17	1.17	1.20	1.18
Item 14	1.22	1.30	1.29	1.30	1.29
Becker05					
Item 3	0.89	0.86	0.86	0.85	0.86
Item 4	1.38	1.30	1.29	1.25	1.27
Item 5	0.72	0.71	0.77	0.79	0.78
Item 6	1.08	1.15	1.11	1.12	1.12
Item 7	0.53	0.60	0.57	0.58	0.57
Item 8	1.46	1.34	1.33	1.32	1.32
Item 10	0.73	0.79	0.78	0.78	0.78
Item 11	1.31	1.33	1.29	1.30	1.30
Item 13	1.17	1.17	1.17	1.20	1.18
Item 14	1.28	1.30	1.29	1.30	1.29
Becker06					
Item 3	0.89	0.86	0.86	0.85	0.86
Item 4	1.51	1.30	1.29	1.25	1.27
Item 5	0.44	0.71	0.43*	0.43*	0.43*
Item 6	1.46	1.15	1.44*	1.44*	1.44*
Item 7	0.50	0.60	0.57	0.58	0.57
Item 8	1.34	1.34	1.33	1.32	1.32
Item 10	0.65	0.79	0.78	0.78	0.78
Item 11	1.29	1.33	1.29	1.30	1.30
Item 13	1.19	1.17	1.17	1.20	1.18
Item 14	1.34	1.30	1.29	1.30	1.29
Becker08					
Item 3	0.89	0.86	0.86	0.85	0.86
Item 4	1.34	1.30	1.29	1.25	1.27
Item 5	0.81	0.71	0.77	0.79	0.78
Item 6	1.16	1.15	1.11	1.12	1.12
Item 7	0.70	0.60	0.57	0.58	0.57

(continued)

Item 8	1.44	1.34	1.33	1.32	1.32
Item 10	0.95	0.79	0.78	0.78	0.78
Item 11	1.46	1.33	1.29	1.30	1.30
Item 13	1.25	1.17	1.17	1.20	1.18
Item 14	1.47	1.30	1.29	1.30	1.29

Note. * denotes freely estimated parameters in partial metric and later models

Table 13

Item Intercepts for the BSQ Across Models

	Model				
	Configural	Metric	Partial Metric	Scalar	Partial Scalar
Stice01					
Item 3	1.74	1.74	1.74	1.97	1.72**
Item 4	3.38	3.38	3.38	3.57	3.52
Item 5	2.29	2.29	2.29	2.09	2.16
Item 6	3.65	3.65	3.65	3.61	3.59
Item 7	1.93	1.93	1.93	1.85*	1.91*
Item 8	2.84	2.84	2.84	2.96	2.93
Item 10	1.81	1.81	1.81	1.95	1.93
Item 11	3.46	3.46	3.46	3.36	3.42
Item 13	3.15	3.15	3.15	2.89	3.15**
Item 14	3.42	3.42	3.42	3.36	3.33
Becker05					
Item 3	1.74	1.74	1.74	1.97	1.98
Item 4	3.49	3.48	3.48	3.57	3.82**
Item 5	1.77	1.78	1.79	2.09	2.00**
Item 6	3.17	3.19	3.19	3.61	3.59
Item 7	1.45	1.47	1.46	1.68	1.65
Item 8	2.63	2.61	2.61	2.96	2.93
Item 10	1.67	1.69	1.68	1.95	1.93
Item 11	3.05	3.07	3.06	3.36	3.42
Item 13	2.40	2.40	2.41	2.89	2.74**
Item 14	3.04	3.05	3.05	3.36	3.33
Becker06					
Item 3	1.74	1.74	1.74	1.97	1.98
Item 4	3.44	3.39	3.39	3.57	3.52
Item 5	1.65	1.74	1.65	1.77*	1.76*
Item 6	3.41	3.33	3.42	3.79*	3.78*
Item 7	1.50	1.53	1.52	1.68	1.65
Item 8	2.59	2.60	2.60	2.96	2.93
Item 10	1.75	1.80	1.80	1.95	1.93
Item 11	3.05	3.07	3.06	3.36	3.42

(continued)

Item 13	2.60	2.60	2.60	2.89	2.86
Item 14	2.92	2.92	2.92	3.36	3.33
Becker08					
Item 3	1.74	1.74	1.74	1.97	1.98
Item 4	2.99	2.99	2.99	3.57	3.38**
Item 5	1.88	1.88	1.88	2.09	2.16
Item 6	3.29	3.29	3.29	3.61	3.59
Item 7	1.51	1.51	1.51	1.68	1.65
Item 8	2.54	2.54	2.54	2.96	2.93
Item 10	1.73	1.73	1.73	1.95	1.93
Item 11	2.82	2.82	2.82	3.36	3.21**
Item 13	2.51	2.51	2.51	2.89	2.86
Item 14	2.87	2.87	2.87	3.36	3.33

Note. * denotes parameters freed from metric model; ** denotes parameters freed in partial scalar model

A scalar invariance model also showed good fit, where item loadings and means were constrained to be equal across all groups, except the loadings and corresponding intercepts released in the partial metric model. The LRT test indicated significantly degraded fit between the models, $\chi^2(24) = 67.50$, $p < .001$. Based on expected parameter change estimated by modification indices, seven item intercepts were released: items 3 ("Have you felt so bad about your shape that you cried?") and 13 ("Have you avoided situations where people could see your body (e.g., swimming pools/communal changing rooms)?") in Stice 2001; items 4 ("Has being with thin women made you feel self-conscious about your shape?"), 5 ("Has thinking about your shape interfered with your ability to concentrate (e.g., while watching television)?"), and 13 in Becker 2005; and items 4 and 11 ("Has seeing your reflection (e.g., in the mirror) made you feel bad about your shape?") in Becker 2008 (see Table 13). Similar to the other measures, the raw item differences were fairly small, as can be seen in Figure 9. The revised model indicated good fit and the LRT test was no longer significant, $\chi^2(17) = 23.26$, $p = .14$, indicating partial scalar invariance across studies.

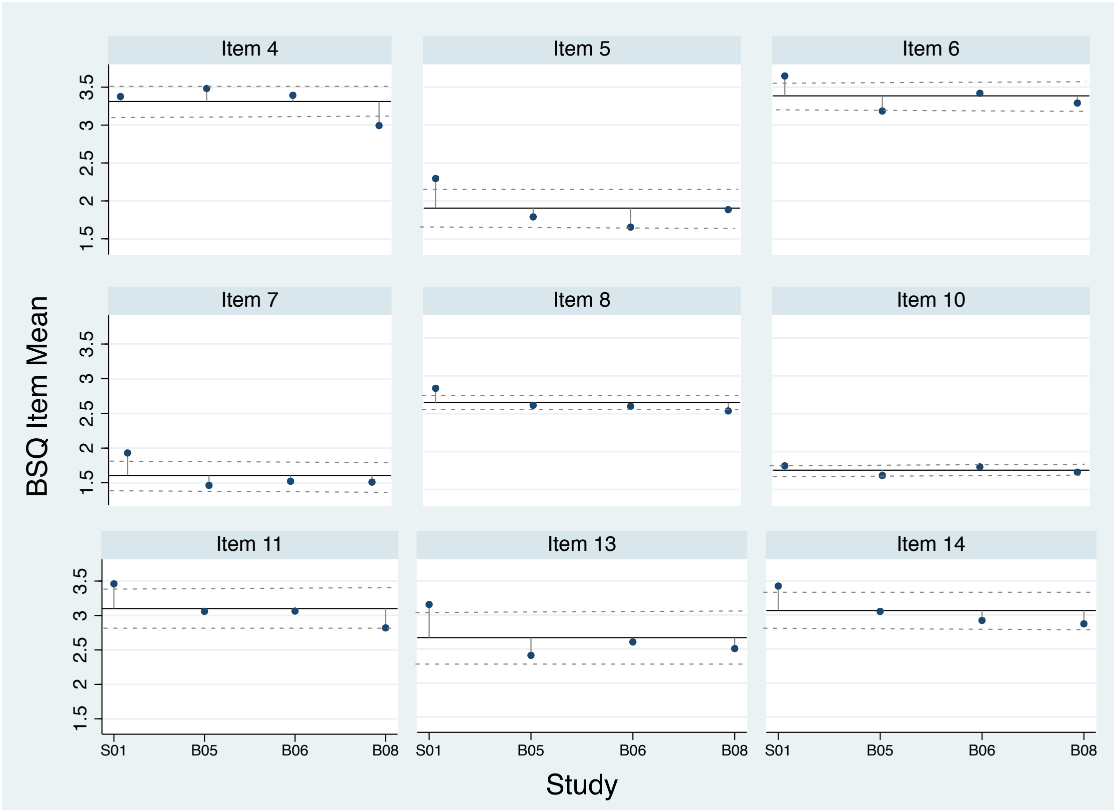

Figure 9. BSQ item means by study shown around the item mean across studies. Dashed lines represent one standard deviation above and below each item mean.

Age Effects

In order to examine the effects of age on factor scores, age was included as a predictor of each latent construct. Analyses indicated that age was a significant predictor of the latent variable for the BDS, with every one year increase in age predicting a 0.03 decrease in BDS score across studies ($p < .05$). Age was also a significant predictor of the latent value of the BSQ, where every one year increase in age predicted a 0.08 decrease in BSQ score ($p < .05$). The effect of age approached significance as a predictor of the IBSS-R score ($p = .06$), where for every one year increase in age, IBSS-R score increased by .03 across studies. Age effects within individual studies can be seen in Table 14.

Table 14

Unstandardized Regression Coefficients for Age Within Studies

		IBSS-R		BDS		BSQ	
Study	N	B	SE	B	SE	B	SE
Stice 2001	87	0.04	0.06	-0.01	0.06	0.00	0.05
Stice 2002	147	0.16**	0.06	0.14*	0.05	-	-
Stice 2006	478	0.20**	0.04	-0.06	0.03	-	-
Stice 2009	303	-0.05	0.06	-0.10*	0.05	-	-
McMillan 2011	123	-0.04	0.03	-0.03	0.02	-	-
Becker 2005	146	-0.16	0.12	-	-	-0.27**	0.08
Becker 2006	74	0.06	0.24	-	-	0.09	0.15
Becker 2008	164	-0.29	0.16	-	-	-0.09	0.11
Becker 2010	101	-0.26	0.18	-	-	-	-

* $p < .05$ ** $p < .001$

Factor Scores

I used the finalized partial scalar invariance models to calculate factor scores. I used bivariate correlations as a preliminary examination of the differences in factor scores using the partial scalar models as compared to each participant's original raw score average. Correlations for each measure ranged between .97 and .99 ($p < .001$) and can be seen in Figure 10.

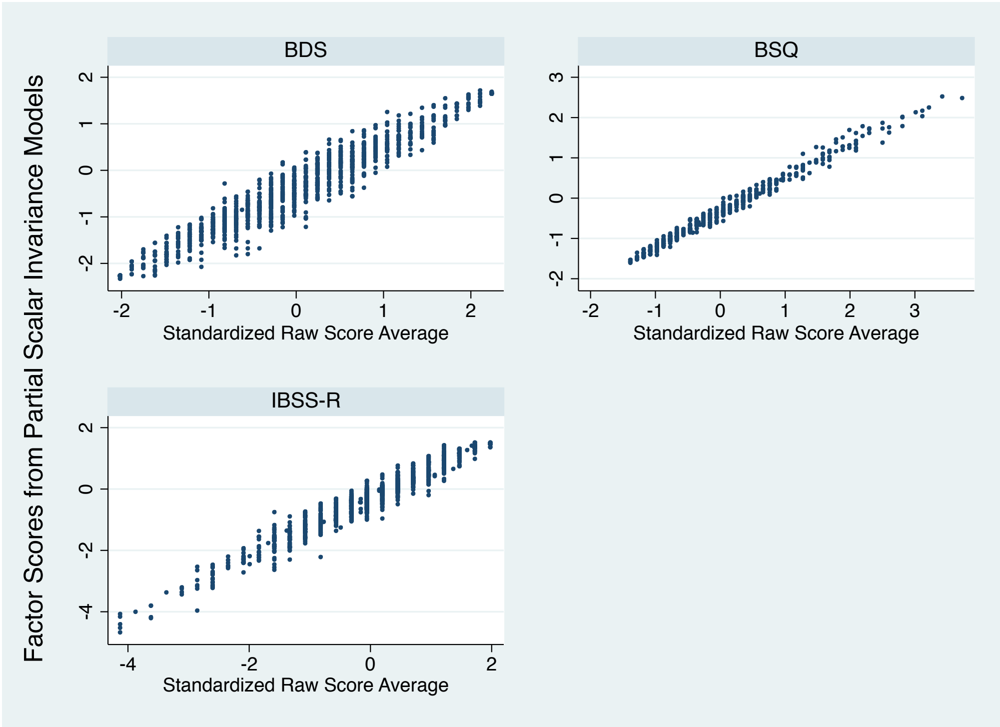

Figure 10. Factor scores calculated from the finalized invariance models correlated with raw score averages for each measure.

Discussion

The purpose of this project was to establish measurement invariance of three measures across 10 heterogeneous data sets focused on dissonance-based interventions for eating disorders. I was able to accomplish my three overall research aims, with some modifications to the originally proposed methods. First, I gathered and organized the data from 10 studies. I then identified the two key constructs of thin ideal internalization and body dissatisfaction. Thin ideal internalization was measured using the Ideal Body Stereotype Scale-Revised (IBSS-R) and was used in all studies. Body dissatisfaction was assessed using two different measures, the Satisfaction and Dissatisfaction with Body Parts Scale (BDS) used in all six of the Stice studies, and the Body Shape Questionnaire (BSQ) used in two of the early Stice studies and three of the Becker studies. After identifying the constructs of thin ideal internalization and body dissatisfaction, I established partial measurement invariance of each measure across studies.

Age Effects

After establishing partial scalar measurement invariance, I examined the effects of age on the latent variable for each measure. Analyses indicated that age effects were inconsistent within individual studies for all measures, but each construct showed a small effect when estimated across studies. Results suggested a small but significant decrease in body dissatisfaction, as assessed by the BDS and the BSQ, with increasing age. In contrast, results suggested a slight increase in thin ideal internalization with increasing age, although this effect was not strictly statistically significant (p = .06). Examining age effects across the aggregated, heterogeneous data sets allowed for examining effects over a wider age range. These findings may help future researchers study and incorporate age characteristics into interventions to increase effectiveness.

Implications

One of the clearest implications of using IDA is the increased flexibility in examining measures and the opportunity for refining measures to better assess target populations. In this study, one example of this was items 7 and 8 of the IBSS-R. I dropped these items in initial analyses given their consistently lower loadings in the studies with younger participants. It appears that the words "curvy" and "shapely" were associated with the thin ideal much less for younger girls than college-aged women. This finding may be due to less familiarity or exposure to those descriptors, a different interpretation of those terms or possibly a different perceived relationship between those words and overall body shape for younger girls. Whatever the reason, comparing the strength of loadings made this difference clear and may warrant adjustments to the measure in continued research on thin ideal internalization.

An IDA framework emphasizes a more deliberate and rigorous approach for measurement in study design and evaluation. Researchers considering combining studies may be able to anticipate areas of measurement difficulties due to age or other significant population characteristics and make needed adjustments while in the planning stages of their research. However, IDA also provides valuable information regarding measurement after data is collected. Examining measurement across studies can highlight areas of strain in measures, indicating potential new findings or areas needing revision.

Future Directions

Clinical significance. One critical question that needs further examination within the development of IDA is the clinical significance of establishing invariance when aggregating data across studies. In a study like this one, where participant-level data is combined across 10 randomized trials conducted by different researchers, some degree of measurement non-invariance is inevitable. However, the impact and importance of the invariance of loadings, intercepts and residuals is not clearly understood. As noted in the results above, adjustments made to the metric and scalar models to achieve partial invariance often reflected fairly small differences in raw means or factor loadings. Some of these differences may not be clinically significant or relevant in treatment settings. Future research needs to explore ways to examine the clinical significance of differences between groups to better understand the impact of invariance on overall treatment effects. In addition to future IDA studies combining randomized trials, simulation studies examining bias in treatment effect based on different degrees of invariance may also aid in this goal.

Examining and addressing questions of clinical significance will likely require some changes to the established process of IDA. For example, although increased sample sizes in combined data allow for increased precision and flexibility in analyses, they may also increase the sensitivity of statistical significance tests such as likelihood-ratio tests. The small raw item mean differences in significant LRT tests and high correlations between the adjusted factor scores and the original scale scores raise questions regarding the efficiency of using LRT tests as the primary mechanism to test differences between models in IDA procedures. Future research should explore using effect sizes rather than LRT tests as a more efficient method to identify and address clinically meaningful differences between groups. Future research should also continue developing methods established to analyze a larger number of studies in IDA, such as random effects modeling.

Research collaboration. This study suggests that IDA is a feasible practice across a larger number of studies than previously examined in the behavioral science literature. The largest difficulty regarding chaining heterogeneous measures across one construct is likely an avoidable problem with coordinated research efforts. Given the significant benefits of IDA, including cost-effective ways to replicate, generalize and expand research findings and build a more cumulative science, we need more collaborative research networks in the field. Researchers can and should continue exploring unique research questions using heterogeneous methods, but some degree of collaboration will make it possible to then aggregate data and increase the breadth, depth, and impact of findings than is possible in any single study. Hofer and Piccinin (2009) suggest one possible model of research collaboration, involving establishing networks of researchers to help facilitate and coordinate research in the targeted area through the proposal, design, analysis and publication processes. It is likely that different approaches to collaboration will be needed depending on the specific topics and research designs. However, any model will likely need to include more coordinated communication between researchers at every stage of research, more data sharing and access to study-level variables (e.g., protocols), and more shared authorship agreements.

Conclusion

In conclusion, this study provided valuable information in the continued development of IDA. To this point there has not been a rigorous, systematic method in place to allow researchers to aggregate item-level data while ensuring measurement invariance, which is a necessary prerequisite for meaningfully interpreting results. Integrative Data Analysis is a proposed method for addressing that deficit and allowing researchers to capitalize on the benefits of pooled participant-level data. Those benefits include, but are not limited to, exploration of original questions with a much larger sample, examination of new hypotheses, more appropriate statistical analyses, broader assessment of constructs and replicability of results without conducting additional individual, costly trials. This project was specifically designed to address the first steps of IDA and establishing measurement invariance of thin ideal internalization and body dissatisfaction across study in 10 heterogeneous eating disorder prevention data sets.

Establishing measurement invariance across the dissonance-based intervention studies allowed for the examination of age effects across a wider range of ages and also provided a foundation for examining treatment effects across all studies as part of future research. This project provides a foundation and resource for other researchers interested in collaboration and data aggregation, especially those who are interested in combining more studies than previously examined and with more disparate methodologies. Particularly at a time where funding is limited but the capacity for data sharing and global collaboration are more advanced than ever, continued efforts to develop IDA procedures are an important step in making psychology a truly cumulative and advancing science.

References

Ackard, D. M., Fulkerson, J. A., & Neumark-Sztainer, D. (2007). Prevalence and utility of DSM-IV eating disorder diagnostic criteria among youth. *International Journal of Eating Disorders*, *40*, 409-417. doi: 10.1002/eat.20389

Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods*, *14*, 101-125. doi: 10.1037/a0015583

Becker, C. B., Bull, S., Schaumberg, K., Cauble, A., & Franco, A. (2008). Effectiveness of peer-led eating disorders prevention: A replication trial. *Journal of Consulting and Clinical Psychology*, *76*, 347-354. doi: 10.1037/0022-006X.76.2.347

Becker, C., Smith, L., & Ciao, A. (2005). Reducing eating disorder risk factors in sorority members: A randomized trial. *Behavior Therapy*, *36*, 245-253. doi: 10.1016/S0005-7894(05)80073-5

Becker, C. B., Smith, L. M., & Ciao, A. C. (2006). Peer-facilitated eating disorder prevention: A randomized effectiveness trial of cognitive dissonance and media advocacy. *Journal of Counseling Psychology*, *53*, 550-555. doi: 10.1037/0022-0167.53.4.550

Becker, C. B., Wilson, C., Williams, A., Kelly, M., McDaniel, L., & Elmquist, J. (2010). Peer-facilitated cognitive dissonance versus healthy weight eating disorders prevention: A randomized comparison. *Body Image*, *7*, 280-288. doi: 10.1016/j.bodyim.2010.06.004

Berg, K. C., Frazier, P., & Sherr, L. (2009). Change in eating disorder attitudes and behavior in college women: Prevalence and predictors. *Eating Behaviors*, *10*, 137-142. doi: 10.1016/j.eatbeh.2009.03.003

Berlin, J. A., Santanna, J., Schmid, C. H., Szczech, L. A., & Felman, H. I. (2002). Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: Ecological bias rears its ugly head. *Statistics in Medicine*, *21*, 371-387. doi: 10.1002/sim.1023

Berscheid, E., Walster, E., & Bohrnstedt, G. (1973). The happy American body: A survey report. *Psychology Today*, *7*(6), 119-131.

Brown, T. A. (2006). *Confirmatory factor analysis: For applied research.* New York, NY: The Guilford Press.

Byrne, B., Shavelson, R., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*, 456-466. doi: 10.1037//0033-2909.105.3.456

Cialdini, R. B., & Fultz, J. (1990). Interpreting the negative mood-helping literature via "mega"-analysis: A contrary view. *Psychological Bulletin*, *107*, 210-214. doi: 10.1037/0033-2909.107.2.210

Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods*, *14*, 165-176. doi: 10.1037/a0015565

Cooper, P. J., Taylor, M. J., Cooper, Z., & Fairburn, C. G. (1987). The development and validation of the Body Shape Questionnaire. *International Journal of Eating Disorders*, *6*, 485-494. doi: 10.1002/1098-108X(198707)6:4<485::AID-EAT2260060405>3.0.CO;2-O

Curran, P. J., & Hussong, A. M. (2009). Integrative Data Analysis: The simultaneous analysis of multiple data sets. *Psychological Methods*, *14*, 81-100. doi: 10.1037/a0015914

Curran, P. J., Hussong, A. M., Cai, L., Huang, W., Chassin, L., Sher, K. J., & Zucker, R. A. (2008). Pooling data from multiple longitudinal studies: The role of item response theory in integrative data analysis. *Developmental Psychology*, *44*, 365-380. doi: 10.1037/0012-1649.44.2.365

de Maat, S., Dekker, J., Schoevers, R., van Aalst, G., Gijsbers-van Wijk, C., Hendriksen, M., … de Jonghe, F. (2008). Short Psychodynamic Supportive Psychotherapy, antidepressants, and their combination in the treatment of major depression: A mega-analysis based on three randomized clinical trials. *Depression and Anxiety*, *25*, 565-574. doi: 10.1002 /da.20305

DeRubeis, R. J., Gelfand, L. A., Zang, T. Z., & Simons, A. D. (1999). Medications versus cognitive behavior therapy for severely depressed outpatients: Mega-analysis of four randomized comparisons. *American Journal of Psychiatry*, *156*, 1007-1013. Retrieved from http://ajp.psychiatryonline.org/

Dowson, J., & Henderson, L. (2001). The validity of a short version of the Body Shape Questionnaire. *Psychiatry Research*, *102*, 263-271. doi: 10.1016/S0165-1781(01)00254-2

Fairburn, C. G., Cooper, Z., Bohn, K., O'Connor, M. E., Doll, H. A., & Palmer, R. L. (2007). The severity and status of eating disorder NOS: Implications for DSM-V. *Behaviour Research and Therapy*, *45*, 1705-1715. doi: 10.1016/j.brat.2007.01.010

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Fournier, J. C., DeRubeis, R. J., Hollon, S. D., Dimidjian, S., Amsterdam, J. D., Shelton, R. C., & Fawcett, J. (2010). Antidepressant drug effects and depression severity: A patient-level meta-analysis. *Journal of the American Medical Association*, *303*, 47-53. doi: 10.1001/ jama.2009.1943

Freeman, C. D., & Strayer, A. H. (1996). Mega-analysis of meta-analysis: An examination of meta-analysis with an emphasis on once-daily aminoglycoside comparative trials. *Pharmacotherapy*, *16*, 1093-1102. Retrieved from: http://pharmacotherapyjournal .org/loi/phco

Green, M., Scott, N., Diyandova, I., Gasser, C., & Pederson, E. (2005). Eating disorder prevention: An experimental comparison of high level dissonance, low level dissonance, and no-treatment control. *Eating Disorders*, *13*, 157-169. doi: 10.1080/ 10640260590918955

Hofer, S. M., & Piccinin, A. M. (2009). Integrative Data Analysis through coordination of measurement and analysis protocol across independent longitudinal studies. *Psychological Methods*, *14*, 150-164. doi: 10.1037/a0015566

Hudson, J. I., Hiripi, E., Pope, H. G., & Kessler, R. C. (2007). The prevalence and correlates of eating disorders in the National Comorbidity Survey Replication. *Biological Psychiatry*, *61*, 348-358. doi: 10.1016/j.biopsych.2006.03.040

Hussong, A. M., Bauer, D., Huang, W., Chassin, L., Sher, K. J., & Zucker, R. A. (2008). Characterizing the life stressors of children of alcoholic parents. *Journal of Family Psychology*, *22*, 819-832. doi: 10.1037/a0013704

Hussong, A. M., Cai, L., Curran, P. J., Flora, D. B., Chassin, L. A., & Zucker, R A. (2008). Disaggregating the distal, proximal, and time-varying effects of parent alcoholism on children's internalizing symptoms. *Journal of Abnormal Child Psychology*, *36*, 335-346. doi: 10.1007/s10802-007-9181-9

Hussong, A. M., Flora, D. B., Curran, P. J., Chassin, L. A., & Zucker, R. A. (2008). Defining risk heterogeneity for internalizing symptoms among children of alcoholic parents. *Development and Psychopathology*, *20*, 165-193. doi: 10.1016/j.bbi.2008.05.010

Hussong, A. M., Wirth, R. J., Edwards, M. C., Curran, P. J., Chassin, L. A., & Zucker, R. A. (2007). Externalizing symptoms among children of alcoholic parents: Entry points for an antisocial pathway to alcoholism. *Journal of Abnormal Psychology*, *116*, 529-542. doi: 10.1037/0021-843X.116.3.529

Kaplow, J. B., Curran, P. J., Dodge, K. A., & The Conduct Problems Prevention Research Group. (2002). Child, parent, and peer predictors of early-onset substance use: A multisite longitudinal study. *Journal of Abnormal Child Psychology*, *30*, 199-216. doi: 10.1023/A:1015183927979

Keski-Rahkonen, A., Hoek, H. W., Linna, M. S., Raevuori, A., Sihvola, E.,…Kaprio, J. (2009). Incidence and outcomes of bulimia nervosa: A nationwide population-based study. *Psychological Medicine*, *39,* 823-831. doi: 10.1017/S0033291708003942

Lambert, P. C., Sutton, A. J., Abrams, K. R., & Jones, D. R. (2002). A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *Journal of Clinical Epidemiology*, *55*, 86-94. doi: 10.1016/S0895-4356(01)00414-0

Matusek, J., Wendt, S., & Wiseman, C. (2004). Dissonance thin-ideal and didactic healthy behavior eating disorder prevention programs: Results from a controlled trial. *International Journal of Eating Disorders*, *36*, 376-388. doi: 10.1002/eat.20059

McMillan, W., Stice, E., & Rohde, P. (2011). High- and low-level dissonance-based eating disorder prevention programs with young women with body image concerns: An experimental trial. *Journal of Consulting and Clinical Psychology*, *79*, 129-134. doi: 10.1037/a0022143

Mitchell, K. S., Mazzeo, S. E., Rausch, S. M., & Cooke, K. L. (2007). Innovative interventions for disordered eating: Evaluating dissonance-based and yoga interventions. *International Journal of Eating Disorders*, *40*, 120-128. doi:10.1002/eat.20282

Pyle, R. L., Neuman, P. A., Halvorson, P. A., & Mitchell, J. E. (1991). An ongoing cross-sectional study of the prevalence of eating disorders in freshman college students. *International Journal of Eating Disorders*, *10*, 667-677. doi: 10.1002/1098-108X(199111)10:6<667::AID-EAT2260100605>3.0.CO;2-K

Roehrig, M., Thompson, J. K., Brannick, M., & van den Berg, P. (2006). Dissonance-based eating disorder prevention program: A preliminary dismantling investigation. *International Journal of Eating Disorders, 39,* 1-10. doi: 10.1002/eat.20217

Rosen, J. C., Jones, A., Ramirez, E., & Waxman, S. (1996). Body Shape Questionnaire: Studies of validity and reliability. *International Journal of Eating Disorders*, *20*, 315-319. doi: 10.1002/(SICI)1098-108X(199611)20:3<315::AID-EAT11>3.0.CO;2-Z

Seretti, A., Cusin, C., Rausch, J. L., Bondy, B., & Smeraldi, E. (2006). Pooling pharmacogenetic studies on the serotonin transporter gene: A mega-analysis. *Psychiatry Research*, *145*, 61-65. doi: 10.1016/j.psychres.2005.09.020

Sternberg, K. J., Baradaran, L. P., Abbott, C. B., Lamb, M. E., & Guterman, E. (2006). Type of violence, age, and gender differences in the effects of family violence on children's behavior problems: A mega-analysis. *Developmental Review*, *26*, 89-112. doi: 10.1016/j.dr.2005.12.001

Stewart, L., & Clarke, M. (1995). Practical methodology of meta-analyses (overviews) using updated individual patient data. *Statistics in Medicine*, *14*, 2057-2079. doi: 10.1002/sim.4780141902

Stice, E, & Agras, W. (1998). Predicting onset and cessation of bulimic behaviors during adolescence: A longitudinal grouping analysis. *Behavior Therapy*, *29*, 257-276. doi: 10.1016/S0005-7894(98)80006-3

Stice, E., Chase, A., Stormer, S., & Appel, A. (2001). A randomized trial of a dissonance-based eating disorder prevention program. *The International Journal of Eating Disorders*, *29*, 247-262. doi: 10.1002/eat.1016

Stice, E., Fisher, M., & Martinez, E. (2004). Eating Disorder Diagnostic Scale: Additional evidence of reliability and validity. *Psychological Assessment*, *16*, 60-71. doi: 10.1037 /1040-3590.16.1.60

Stice, E., Marti, C. N., Shaw, H., & Jaconis, M. (2010). An 8-year longitudinal study of the natural history of threshold, subthreshold and partial eating disorders from a community sample of adolescents. *Journal of Abnormal Psychology*, *118*, 587-597. doi: 10.1037 /a0016481

Stice, E., Marti, C. N., Spoor, S., Presnell, K., & Shaw, H. (2008). Dissonance and healthy weight eating disorder prevention programs: Long-term effects from a randomized efficacy trial. *Journal of Consulting and Clinical Psychology*, *76*, 329-340. doi: 10.1037/0022-006X.76.2.329

Stice, E., Mazotti, L., Weibel, D., & Agras, W. S. (2000). Dissonance prevention program decreases thin-ideal internalization, body dissatisfaction, dieting, negative affect, and bulimic symptoms: A preliminary experiment. *International Journal of Eating Disorders*, *27*, 206 - 217. doi: 10.1002/(SICI)1098-108X(200003)27:2<206::AID-EAT9>3.0.CO;2-D

Stice, E., Nemeroff, C., & Shaw, H. (1996). Test of dual pathway model of bulimia nervosa: Evidence for dietary restraint and affect regulation mechanisms. *Journal of Social and Clinical Psychology*, *15*, 340-363. Retrieved from www.guilford.com/pr/jnsc.htm

Stice, E., Presnell, K., Gau, J. & Shaw, H. (2007) Testing mediators of intervention effects in randomized controlled trials: An evaluation of two eating disorder prevention programs. *Journal of Consulting and Clinical Psychology*, *75*, 20-32. doi: 10.1037/0022-006X.75.1.20

Stice, E., Rohde, P., Gau, J., & Shaw, H. (2009). An effectiveness trial of a dissonance-based eating disorder prevention program for high-risk adolescent girls. *Journal of Consulting and Clinical Psychology*, *77*, 825-834. doi: 10.1037/a0016132

Stice, E., & Shaw, H. (2002). Role of body dissatisfaction in the onset and maintenance of eating pathology: A synthesis of research findings. *Journal of Psychosomatic Research*, *53*, 985-993. doi: 10.1016/S0022-3999(02)00488-9

Stice, E., & Shaw, H. (2004). Eating disorder prevention programs: A meta-analytic review. *Psychological Bulletin*, *130*, 206-227. doi: 10.1037/0033-2909.130.2.206

Stice, E., Shaw, H., Becker, C. B., & Rohde, P. (2008). Dissonance-based interventions for the prevention of eating disorders: Using persuasion principles to promote health. *Prevention Science*, *9*, 114-128. doi: 10.1007/s11121-008-0093-x

Stice, E., Shaw, H., Burton, E., & Wade, E. (2006). Dissonance and healthy weight eating disorder prevention programs: A randomized efficacy trial. *Journal of Consulting and Clinical Psychology*, *74*, 263-275. doi: 10.1016/j.bbi.2008.05.010

Stice, E., Shaw, H., & Marti, C. (2007). A meta-analytic review of eating disorder prevention programs: Encouraging findings. *Annual Review of Clinical Psychology*, *3*, 207-231. doi: 10.1146/annurev.clinpsy.3.022806.091447

Stice, E., Trost, A., & Chase, A. (2003). Healthy weight control and dissonance-based eating disorder prevention programs: Results from a controlled trial. *International Journal of Eating Disorders*, *33*, 10-21. doi: 10.1002/eat.10109

Stice, E., Ziemba, C., Margolis, J., & Flick, P. (1996). The dual pathway model differentiates bulimics, subclinical bulimics, and controls: Testing the continuity hypothesis. *Behavior Therapy*, *27*, 531-549. doi: 10.1016/S0005-7894(96)80042-6

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73-140). Mahwah, NJ: Erlbaum.

Touchette, E., Henegar, A., Godart, N. T., Pryor, L., Falissard, B., Tremblay, R. E., & Cote, S. M. (2011). Subclinical eating disorders and their comorbidity with mood and anxiety disorders in adolescent girls. *Psychiatry Research*, *185*, 185-192. doi: 10.1016/j.psychres .2010.04.005