An Item Response Theory Analysis of Emotional Expressivity Measures

Melissa K. Matheys
Illinois State University, mmathey@ilstu.edu

Follow this and additional works at: https://ir.library.illinoisstate.edu/etd
Part of the Quantitative Psychology Commons

Recommended Citation
https://ir.library.illinoisstate.edu/etd/623

This Thesis and Dissertation is brought to you for free and open access by ISU ReD: Research and eData. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of ISU ReD: Research and eData. For more information, please contact ISURed@ilstu.edu.
AN ITEM RESPONSE THEORY ANALYSIS
OF EMOTIONAL EXPRESSIVITY
MEASURES

Melissa K. Matheys

71 Pages

The understanding of psychometric properties of emotional expressivity measures is often limited to rudimentary statistics. The present study investigated, in depth, the psychometric properties of four measures of emotional expressivity (the Self-Concealment Scale, the Distress Disclosure Index, the Emotional Expressivity Scale, and the Emotion Regulation Questionnaire – Suppression subscale). Item response theory (IRT) analyses suggested that recoded versions of the measures’ Likert scales may be beneficial in practice as well as the deletion of some measures’ items. IRT analyses of the four measures of emotional expressivity provide more information than classical test theory statistics about how to implement these measures best.

KEYWORDS: Distress Disclosure, Emotional Expressivity, Item Response Theory, Self-Concealment, Suppression
AN ITEM RESPONSE THEORY ANALYSIS
OF EMOTIONAL EXPRESSIVITY
MEASURES

MELISSA K. MATHEYS

A Thesis Submitted in Partial
Fulfillment of the Requirements
for the Degree of

MASTER OF SCIENCE

Department of Psychology

ILLINOIS STATE UNIVERSITY

2016
AN ITEM RESPONSE THEORY ANALYSIS
OF EMOTIONAL EXPRESSIVITY
MEASURES

MELISSA K. MATHEYS

COMMITTEE MEMBERS:
Jeffrey H. Kahn, Chair
Matthew S. Hesson-McInnis
CONTENTS

<table>
<thead>
<tr>
<th>CONTENTS</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>TABLES</td>
<td>iii</td>
</tr>
<tr>
<td>FIGURES</td>
<td>iv</td>
</tr>
</tbody>
</table>

CHAPTER

I. THE PROBLEM AND ITS BACKGROUND 1

Statement of the Problem 1

II. REVIEW OF THE LITERATURE 6

Emotional Expressivity 6

Theoretical Background on Emotions 6
Emotional Expressivity Defined 8
Measures of Emotional Expressivity 9

Self-concealment and the Self-Concealment Scale 9
Distress disclosure and the Distress Disclosure Index 12
Emotional expressivity and the Emotional Expressivity Scale 14
Suppression and the Emotion Regulation Questionnaire – Suppression subscale 16

Scales’ Relation to the Model of Emotion 18
Moving Forward 19

Item Response Theory 20
Purpose and Research Questions 26

III. METHOD 28

Participants 28
Measures 28
## TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Means, Standard Deviations, and Reliabilities for Study Scales</td>
<td>33</td>
</tr>
<tr>
<td>2.</td>
<td>Fit Indices Among Study Scales</td>
<td>35</td>
</tr>
<tr>
<td>3.</td>
<td>Graded Response Model Item Parameter Estimates and Item-Fit Statistics for 10-Item Five-Category Self-Concealment Scale</td>
<td>38</td>
</tr>
<tr>
<td>4.</td>
<td>Graded Response Model Item Parameter Estimates and Item-Fit Statistics for 12-Item Recoded Four-Category Distress Disclosure Index</td>
<td>42</td>
</tr>
<tr>
<td>5.</td>
<td>Graded Response Model Item Parameter Estimates and Item-Fit Statistics for 17-Item Recoded Five-Category Emotional Expressivity Scale</td>
<td>46</td>
</tr>
<tr>
<td>6.</td>
<td>Graded Response Model Item Parameter Estimates and Item-Fit Statistics for 4-Item Recoded Six-Category Emotion Regulation Questionnaire – Suppression Subscale</td>
<td>51</td>
</tr>
</tbody>
</table>
## FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Example of an Item Characteristic Curve Ranging from Theta of -3 to Theta of +3</td>
<td>21</td>
</tr>
<tr>
<td>2.</td>
<td>Item Characteristic Curves for a 5-Point Likert Scale</td>
<td>23</td>
</tr>
<tr>
<td>3.</td>
<td>Item Information Curves for Each Rating on a 5-Point Likert Scale</td>
<td>25</td>
</tr>
<tr>
<td>4.</td>
<td>Test Information Curve for all 10 Items on a 5-Point Likert Scale</td>
<td>26</td>
</tr>
<tr>
<td>5.</td>
<td>Test Information Curve for 10-Item Self-Concealment Scale</td>
<td>40</td>
</tr>
<tr>
<td>6.</td>
<td>Test Information Curve for 12-Item Distress Disclosure Index</td>
<td>44</td>
</tr>
<tr>
<td>7.</td>
<td>Test Information Curve for 17-Item Emotional Expressivity Scale</td>
<td>49</td>
</tr>
<tr>
<td>8.</td>
<td>Test Information Curve for 4-Item Emotion Regulation Questionnaire – Suppression Subscale</td>
<td>53</td>
</tr>
<tr>
<td>9.</td>
<td>Item Information Curve for Item 1 of the Distress Disclosure Index</td>
<td>58</td>
</tr>
</tbody>
</table>
CHAPTER I
THE PROBLEM AND ITS BACKGROUND

Statement of the Problem

An emotion, whether it is happiness, sadness, anger, or fear, contains several components. In addition to a subjective experience of emotion, emotions are expressed (Gross 1998; Gross & John, 1997). Because different forms of emotion regulation modulate the final experience and expression of one’s emotion, one’s emotional experience and expression do not need to mirror one another (Kennedy-Moore & Watson, 2001).

Emotional expression can be defined as the verbal or nonverbal communication (Kennedy-Moore & Watson, 2001) of positively or negatively valenced information. In this thesis I have combined several definitions of emotional expressivity to form a comprehensive definition. Ultimately, emotional expressivity can be viewed along a continuum from the expression to the non-expression/suppression of information (Kahn & Hessling, 2001; Larson & Chastain, 1990). A thorough investigation of emotional expressivity would explore constructs that address different combinations of these three defining features: high vs. low expressivity, verbal or non-verbal expression, and valence of information. Of interest to this study are the following four instruments: the Self-Concealment Scale (SCS), the Distress Disclosure Index (DDI), the Emotional
Expressivity Scale (EES), and the Emotion Regulation Questionnaire – Suppression subscale (ERQ-S).

The SCS measures self-concealment, defined as the degree to which someone hides information about themselves from others (Larson & Chastain, 1990). The SCS is a measure of the concealment of verbal information. Psychometric properties of scores from the SCS are good, but researchers have reached different conclusions as to the dimensionality of the SCS (Cramer & Barry, 1999). The SCS has a strong positive predictive validity with depression and anxiety. Self-concealment has a negative correlation with help-seeking attitudes, which Cramer and Barry suggest is maladaptive for the therapeutic relationship. Overall, self-concealment is associated with negative health outcomes.

The DDI measures distress disclosure, or disclosure of personally distressing information (Kahn & Hessling, 2001). Low scores on the DDI reflect frequent concealment of information, and high scores reflect frequent disclosure of information. Psychometric properties of DDI scores are strong. The DDI is related to positive psychological outcomes such as social support and positive affect at the time of testing as well as increased self-esteem and life satisfaction at a later time. Research indicates positive health outcomes for those who disclose personally distressing information.

Emotional expressivity, as operationalized by the EES, reflects both positive and negative displays of emotion that are either verbal or nonverbal. Psychometric properties of EES scores are strong. Kring, Smith, and Neale (1994) thoroughly assessed the EES’s convergent validity: the EES has a strong positive correlation with satisfaction with life
and with positive and negative emotions. That is, the EES is a general measure of emotional expression regardless of valence of channel.

Finally, the ERQ-S measures suppression, which is another word for withholding an emotion (Gross & John, 2003). The withheld emotion can be either positive or negative. Psychometric properties for scores from the ERQ-S are acceptable. The ERQ-S is positively correlated with rumination and negatively correlated with negative mood regulation. It appears that suppression is associated with mostly negative outcomes. More specifically, recent research suggests that whether or not suppression is harmful depends on one’s self-regulatory strength.

Despite the adequate psychometric properties of scores from these four measures of emotional expressivity, each scale relies on averaged (or summed) item responses. That is, each scale relies on aggregate measures of the construct (Fraley, Waller, & Brennan, 2000), which means reliability estimates are the same for each person and each item for each scale. Statistics based on classical test theory such as Cronbach’s alpha do not capture as precise of reliability estimates as does item response theory (IRT). In general, IRT analyses indicate whether items or tests suffer from scaling problems (i.e., is the difference between a 1 [strongly disagree] and a 2 [disagree] on a Likert scale the same as the difference between a 2 and a 3 [neither agree nor disagree], etc.). In IRT, researchers assume a latent trait, theta, exists (Embretson & Reise, 2000). The latent trait in this line of research is emotional expressivity. Graphically, theta values range from -3 to +3 on the abscissa in 1 standard deviation increments. Theta values indicate direction and magnitude of the trait.
The two-parameter logistic (2PL) model is the most common way to use IRT to analyze a self-report measure scored on a Likert scale. The 2PL model estimates parameters of item difficulty and item discrimination (which are described in depth in Chapter 2). An IRT analysis of a 2PL model generates a series of graphs that can be used to interpret item and test properties. The item characteristic curve (ICC) shows the relation between the latent trait and the probability of item endorsement. The item information curve (IIC) shows the relation between the latent trait and the probability of selecting a response option on the scale. The test information curve (TIC) shows the relation between the latent trait and overall scale precision.

TICs are the sum of information in IICs across scale items. This feature of IRT analysis builds on classical test theory concepts such as reliability and validity. By constructing IICs and a TIC, IRT analysis overcomes the assumption that scores are reliable for all participants. That is, responses to different items and from different individuals may have different reliability for each item or test. Researchers can also understand how individual differences in the latent trait variable influence the overall understanding gained from item responses (Embretson & Reise, 2000).

The purpose of the current study was to conduct an IRT analysis for the SCS, the DDI, the EES, and the ERQ-S. This purpose was useful to pursue because these scales are widely used in the literature on concealment and disclosure of emotions in everyday life (Greenland, Scourfield, Maxwell, Prior, & Scourfield, 2009), in romantic relationships (Gross & John, 2003), and in therapeutic relationships (Kahn, Achter, & Shambaugh, 2001). The current study was designed to determine the precision of each
scale across latent trait values, which is often assumed to be equal across people and items for each scale.
CHAPTER II
REVIEW OF THE LITERATURE

Emotional Expressivity

Theoretical Background on Emotions

Experiencing and expressing emotions is a natural process that is part of everyday life. For decades, scholars have developed models of emotional expressivity and documented outcomes of different expressivity patterns (Kahn & Hessling, 2001; King, Emmons, & Woodley, 1992; Kring et al., 1994; Larson & Chastain, 1990). Whereas some people express lots of emotions, others express few emotions. Whereas some people express mostly positive emotions, others express mostly negative emotions. An individual’s amount and type of expression fall on some sort of continua. Researchers must now ascertain the intricacies of such models and their assumed continua before implementing them in research.

One of the most popular models of emotional expressivity is Gross and John’s (1997, 1998) model of emotion. The authors argue that emotions begin with emotional cues. More specifically, emotions begin with an evaluation of internal and external emotional cues. Internal emotional cues might include thoughts running through one’s head that a movie is scary. External emotional cues may entail the physical act of going to the movie theater. These evaluations of internal and external emotional cues serve as the input of the model and lead to emotional response tendencies.
Antecedent-focused emotion regulation occurs at the input of the system before the response tendency has ever been generated (Gross, 1998). Another way to think of it is as things people can do before the emotional response is fully activated or has fully changed their behavior. There are four methods of antecedent-focused emotion regulation, which include cognitive reappraisal, attention deployment, situation selection, and situation modification. For the scary movie example, if your emotional cue triggers feelings of fear, you might reappraise a scary scene as having good cinematography or defer your attention from the screen during scary scenes. Each of these modulate the emotional cue that leads to the emotional response tendency.

Gross and John (1997, 1998) define emotional response tendencies as how one reacts to the emotion. These reactions to emotions occur at the behavioral, experiential, and physiological levels. This response tendency means how one reacts behaviorally (e.g., jumping at a scary scene in a movie), how one reacts experientially (i.e., the feeling of doom one gets in a scary movie), and how one reacts physiologically (e.g., increased heart rate when you are afraid) are due to one’s evaluations of one’s emotional cues. Coherence findings between behavioral, experiential, and physiological evaluations, however, are mixed. Mauss, Levenson, McCarter, Wilhelm, and Gross (2005) measured behavioral, experiential, and physiological responses second-by-second during films with amusing and sad content. Results suggest behavior and experience were highly correlated but physiological responses were only moderately correlated with behavior and experience.

On the other hand, response-focused emotion regulation involves modulation at the point of output of the model. Another way to think about this is what people do after
the emotion is experienced. Examples of response-focused emotion regulation include strategies that intensify, diminish, prolong, or curtail ongoing emotional experience, expression, or physiological responding (Gross, 1998). At a scary movie, we might attempt to curtail our feelings of fear and anxiety. Specifically, suppression (i.e., the inhibition of emotion-expressive behavior) is an example of response-focused emotion regulation (Geisler & Schröder-Abé, 2015). Because of individual differences in emotion regulation, emotional expressions may or may not mirror one’s earlier emotional experiences (Kennedy-Moore & Watson, 2001).

Finally, Gross (1998; Gross & John, 1997) states that emotional response tendencies lead to emotional responses. Emotional responses encompass everything from one’s observable reactions to one’s verbal expressions: it is how one reacts to one’s emotions. When a gory scene comes on at the scary movie, observable reactions can range from laughing to getting up and leaving the theater. Verbal expressions might include screaming or cursing at the movie. Overall, Gross and John’s (1997; 2003) model of emotion distinguishes the emotional experience from emotional expression.

**Emotional Expressivity Defined**

Researchers have been investigating the construct of emotional expressivity for years, and there are many different ways of defining this construct. One of the major ways to distinguish among constructs of emotional expressivity is to state whether they measure high expressivity (often called disclosure) or low expressivity (often called concealment) of information (Kahn & Hessling, 2001; Larson & Chastain, 1990). Typically, high expressivity means revealing information, whereas low expressivity means hiding information. A second distinguishing aspect of constructs of emotional
expressivity is whether the expression of information is verbal or non-verbal (Kennedy-Moore & Watson, 2001). Verbal means whether one talks about it (express it in words), and non-verbal refers to behaviors (e.g., smiling, frowning, laughing, and crying). A third distinguishing aspect of emotional expressivity involves the valence of the information, that is, whether the emotional information is positive or negative. Emotional expressivity is also an umbrella term that includes such constructs as self-concealment, distress disclosure, and suppression (Gross & John, 2003; Kahn & Hessling, 2001; Larson & Chastain, 1990). Although related, these constructs have been demonstrated to be unique and worthy of their own measurement scales.

A well-rounded investigation of emotional expressivity constructs would include constructs that measure general expressivity as well as differing degrees of disclosure versus concealment, verbal versus nonverbal expression, and expression of both valences. Fortunately, there exist different instruments to measure these different elements of emotional expressivity: the Self-Concealment Scale (SCS), the Distress Disclosure Index (DDI), the Emotional Expressivity Scale (EES), and the Emotion Regulation Questionnaire – Suppression (ERQ-S) subscale.

**Measures of Emotional Expressivity**

**Self-concealment and the Self-Concealment Scale.** For starters, self-concealment is the degree to which someone actively hides information about themselves from others (Larson & Chastain, 1990). That is, self-concealment is a construct of low expressivity that measures lack of verbal disclosure of negative emotions. An example of verbally withholding a negative emotion would be choosing not to tell someone that you are sad or angry. The Self-Concealment Scale (SCS) was developed as a way to measure
this construct. Low scores indicate low active concealment of negative information, and high scores indicate a high active concealment of negative information. The SCS is a 10-item self-report scale, and its scores have good internal consistency (α = .83 – .87) and good test-retest reliability (.74 for a small group of women after 7 weeks). Split-half reliability was determined to be .79 (Cramer & Barry, 1999).

The SCS has demonstrated good validity: self-concealment has a strong positive correlation with neuroticism, negative affect, depressive symptoms, and anxiety, as well as a strong negative correlation with extraversion and positive affect (Kahn & Hessling, 2001). In addition, self-concealment has a strong positive correlation with depression, anxiety, and physical symptoms; a strong negative correlation with social support from others (Larson & Chastain, 1990); and a negative correlation with subjective well-being (Wang, Qi, & Cui, 2014). Larson and Chastain (1990) state that self-disclosure is often a hallmark of a thriving therapeutic relationship (i.e., between therapist and patient) and that the concealment of traumatic events has negative health outcomes. Research has shown a strong negative correlation between self-concealment and help-seeking attitudes as well as between self-concealment and therapy outcomes (Larson, Chastain, Hoyt, & Ayzenberg, 2015). Therefore, self-concealment is generally associated with negative health outcomes both concurrently and subsequently.

In constructing the Self-Concealment Scale (SCS), Larson and Chastain (1990) conducted a factor analysis to show that self-concealment and self-disclosure are distinct constructs. By comparing the SCS to the established Self-Disclosure Index, the authors concluded that each measure loaded onto its own factor despite a negative correlation between the two (-.27). According to Larson and Chastain (1990), items on the SCS
address three aspects of the construct: “self-reported tendency to keep things to oneself, possession of a personally distressing secret…, and apprehension about the disclosure of concealed personal information.” Yet the authors maintain the unidimensionality of the scale.

Research by Cramer and Barry (1999) suggests problems with the dimensionality of the SCS. The authors additionally noted that the SCS items may load onto two different factors. In Study 1, Cramer and Barry conducted a principal axis factor analysis and extracted two factors, keeping secrets (eigenvalue = 4.43, 3.94 after rotation) and personal concealment (eigenvalue = 1.10, 0.81 after rotation). In Study 2, Cramer and Barry conducted a confirmatory factor analysis on the SCS, and all fit indices except the $\chi^2$ likelihood ratio suggested good fit. The authors note that the two-factor model explained significantly more variance than the one-factor model. Despite these results, Cramer and Barry conclude that the high internal consistency estimates suggest the SCS is a unidimensional scale. This conclusion raises questions as to the dimensionality of the SCS.

Still subsequent research investigated more precisely what self-concealers conceal, as measured by the SCS. Wismeijer (2011) found that self-concealment was positively correlated with keeping secrets of personal inadequacy, sorrows, and worries related to oneself. Because there was no relation between self-concealment and general secrets, these results suggest the SCS measures a subset of private topics about oneself as opposed to keeping secrets in general.
**Distress disclosure and the Distress Disclosure Index.** Self-disclosure, similar to other constructs of emotional expressivity, is the degree to which someone shares information about himself or herself. Sometimes self-disclosure is about personally distressing information. For example, people may choose to share or conceal upsetting thoughts, bad moods, or personal problems. The disclosure of this type of personal information is called distress disclosure (Kahn & Hessling, 2001), and the Distress Disclosure Index (DDI) is a measure of this construct. Barr, Kahn, and Schneider (2008) found high correlations between the DDI and the Emotional Expressivity Scale (.71) and the SCS (-.38). Uysal, Lin, and Knee (2010) also found a strong negative correlation between the DDI and the SCS (-.46). A factor analysis (Barr et al., 2008) indicated the DDI and the EES loaded on a common factor of general expressivity, but the DDI and SCS loaded onto two different factors (i.e., general expressivity and comfort with disclosing, respectively). More research shows a correlation between the DDI and the SCS (-.35) and the Self-Disclosure Index (.43) (Kahn & Hessling, 2001). A confirmatory factor analysis by Kahn and Hessling (2001, Study 2), however, determined that a three-correlated factor model fit the three scales better than a one-factor model. This factor analysis did not quite reach statistical significance, but the results suggest each scale measures a distinct construct. Thus, it is possible to be a high self-discloser but not be a high distress discloser. Kahn, Hucke, Bradley, Glinski, and Malak (2012) also found a high correlation between distress disclosure and expressive suppression, but a confirmatory factor analysis indicated the two constructs do not represent the same trait. While similar to other measures of emotional expressivity at face value, distress
disclosure is a construct of high expressivity involving verbal disclosure of negative information.

Similar to findings by Kahn and Hessling (2001), Kennedy-Moore and Watson (2001) discuss a paradox of distress expression. That is, expression of negative feelings both signifies initial distress as well as a way of coping with that distress. There are three mechanisms by which expression can help one feel better: expression can alleviate distress about distress, expression can facilitate insight, and expression can affect interpersonal relationships in desired ways. Whereas individual differences exist in comfort with expression, researchers agree the benefits of expression take some time to emerge and that expressing positive emotions can additionally help counterbalance distressing emotions.

The DDI is a 12-item bipolar, self-report scale that measures the degree to which someone shares versus conceals unpleasant information about themselves (Kahn & Hessling, 2001). Low scores on the DDI reflect frequent concealment of information, and high scores reflect frequent disclosure of information. Internal consistency of the scores on the DDI is very strong ($\alpha = .92 \text{ to } .95$) (Kahn & Hessling, 2001; Kahn et al., 2012). Test-retest reliability is also strong ($\alpha = .80$) at intervals of 1 week to 4 months. Criterion-related validity is good because the DDI is predictive of actual disclosure inside and outside of the lab, by the self, by a confidant, or by observers. In Kahn and Hessling’s (2001) validation sample in Study 1, the DDI had good convergent validity. That is, the DDI had a strong positive correlation with self-disclosure, social support, and extraversion; a positive correlation with positive affect; and a negative correlation with self-concealment and depressive symptoms. Other research suggests that higher distress
disclosure is positively correlated with private self-consciousness, social support, and femininity, whereas lower distress disclosure is associated with shame (Greenland et al., 2009). These results suggest the DDI is related to positive psychological outcomes at the time of testing. In Study 3 (Kahn & Hessling, 2001), the DDI had predictive validity because it was correlated with subsequent increased self-esteem, life satisfaction, and perceived social support. Contrary to hypotheses, distress disclosure at Time 1 did not predict subsequent depressive symptoms. Overall, research indicates positive health implications for those who disclose personally distressing information.

**Emotional expressivity and the Emotional Expressivity Scale.** Kring et al. (1994), the developers of the Emotional Expressivity Scale (EES), have defined emotional expressivity as the “extent to which people outwardly display their emotions, regardless of valence or channel (i.e., facial, vocal, gestural)” (p. 934). This definition is broad in that it encompasses both positive and negative displays of emotions. In addition, using their definition, emotional expressivity captures both verbal and non-verbal signs of emotion and may be considered a measure of high expressivity. The EES is therefore a general measure of emotional expressiveness. At the time of the EES creation, the only comparable scale was the Emotional Expressivity Questionnaire (EEQ), which has three subscales that measure expression of positive emotion, expression of negative emotion, and expression of intimacy (King et al., 1992). By developing the EES, Kring et al. (1994) hoped to create a more general measure of emotional expression. Although the EES had a strong positive correlation with the EEQ across different samples (.53 and .64), 72% of the variance in each scale was unshared. The authors concluded the EES measures different emotional expressivity than does the EEQ. When compared to other
measures of emotional expressivity, Barr et al. (2008) found evidence of a strong negative correlation between the EES and SCS (-.34). A factor analysis indicated the EES loaded on a general expressivity factor; however, and the SCS loaded on a comfort with disclosing factor.

The EES is a 17-item self-report scale that is unidimensional. Scores range from low emotional expressivity to high emotional expressivity (Kring et al., 1994). In the seminal article on the EES development, scores from the EES were highly reliable (α = .90 - .93) and had strong test-retest reliability (α = .91) after 4 weeks. Kring et al. (1994) assessed the EES’s convergent validity thoroughly through the use of three samples. Undergraduate students had a strong positive correlation with intensity of affect, surgency, and satisfaction with life; a moderate positive correlation with emotional stability; and a strong negative correlation with social anhedonia at the time of testing. The researchers investigated the convergent validity among a sample of upper-level psychology undergraduates: the EES had a strong positive correlation with the EEQ, with positive emotions, with negative emotions, with intimacy, and with intensity of affect. Finally, a sample of introductory psychology undergraduates provided more information on the EES’s convergent validity: the EES had a strong positive correlation with the EEQ, positive emotion, negative emotion, intimacy, affect communication, social closeness, and family expressiveness; the EES only had a moderate positive correlation with affect intensity and satisfaction with life. Overall, the strongest convergent validity was the positive correlations of the EES with intensity of affect, satisfaction with life, and positive and negative emotions. Such validity means that people who scored high on the
EES (i.e., had high emotional expressivity) felt these feelings more intensely, were more satisfied with life, and felt both positive and negative emotions.

Similarly, a study that utilized experience-sampling methodology (ESM) assessed emotional expressivity in daily life. Main effects include a positive correlation between high EES scores and positive affect and between high EES scores and social functioning (Burgin et al., 2012). Looking closer, it appears that people with high EES scores had a stronger relation than people with low EES scores between liking the people they were with and feeling their time with that person was important as well as feeling close to those people. Discriminant validity indicated that there was no relation between the EES and agreeableness, conscientiousness, self-monitoring, social desirability, well-being, stress reaction, and self-esteem across the three samples of undergraduates (Kring et al., 1994). The strong convergent validity supports the EES as a general measure of emotional expression regardless of valence or channel.

**Suppression and the Emotion Regulation Questionnaire-Suppression subscale.** Suppression, sometimes thought of as the opposite of emotional expressivity, reflects how much someone withholds an emotion once it is recognized (Gross & John, 2003). Suppression is a response-focused emotion regulation strategy (Gross, 1998), a topic discussed in detail earlier. It is a measure of low expressivity because it involves actively inhibiting one’s emotions verbally and non-verbally. The withheld information in suppression can be either positive or negative; some examples include keeping a straight face, stifling laughter, or fighting back tears. Gross and John (2003) created the Emotion Regulation Questionnaire – Suppression (ERQ-S) subscale to measure this construct.
The 4-item Emotion Regulation Questionnaire-Suppression (ERQ-S) subscale measures the degree to which people inhibit their responses to negative and positive emotions (Gross & John, 2003). Low scores represent low suppression, and high scores represent high suppression. Psychometric properties for the scores of the ERQ-S are acceptable: internal consistency was .73, and test-retest reliability was .69 after 3 months. Convergent validity exists such that suppression is positively correlated with regulation success, inauthenticity, and rumination, and it is negatively correlated with a venting coping style, attention, clarity, and negative mood regulation. Concurrently, it appears that suppression is associated with mostly negative outcomes. Discriminant validity exists such that there is no correlation between suppression and neuroticism, cognitive ability, SAT performance, nor social desirability concurrently. The ERQ-S is also positively correlated with the SCS (.44) and negatively correlated with the DDI (-.61) (Kahn et al., 2012; Larson et al., 2015). That is, higher scores on the ERQ-S are associated with higher scores (more concealment) on the SCS and lower scores (more concealment) on the DDI.

In addition, the ERQ-S has implications for longer-term well-being. The ERQ-S has a positive correlation with depression and a negative correlation with self-esteem, life satisfaction, optimism, and well-being (Gross & John, 2003). Suppression also has longer-term implications for interpersonal functioning. The ERQ-S was positively associated with avoidance of attachment and negatively associated with social support and sharing positive and negative emotions with others. Gross and John (2003) argued that suppression is taxing both cognitively and socially. The negative health implications support the taxing nature of suppression.
As an example of how the ERQ-S has been used in research, Geisler and Schröder-Abé (2015) investigated whether emotional suppression was beneficial or harmful. Using the limited strength model of self-control, Geisler and Schröder-Abé (2015) stated that acts of control over dominant responses, such as experiencing or expressing an emotion, draw on a resource called self-regulatory strength. Self-regulatory strength has to do with self-control and the idea that one only has so much self-control at one’s disposal. That is, once self-control is used on a particular task (e.g., emotion expression), there is little self-control left for other tasks (i.e., emotion experience). Geisler and Schröder-Abé (2015) used a biological measure called high-frequency heart rate variability to assess the construct of self-regulatory strength such that higher high-frequency heart rate variability scores reflect higher self-regulatory strength. Results suggest that participants with low high-frequency heart rate variability have a positive correlation between expressive suppression and negative affect, whereas participants with high high-frequency heart rate variability have a positive correlation between expressive suppression and partner relationship satisfaction. That is, expressive suppression requires self-control, and the answer to whether or not suppression is harmful lies in one’s self-regulatory strength.

Scales’ Relation to the Model of Emotion

Gross and John’s (1997) aforementioned model of emotional expressivity fits these four constructs neatly. In their aforementioned model, Gross and John (1997) state that emotion occurs when input from the environment or within the person triggers an emotional cue, positive or negative. The emotional cue then creates verbal or nonverbal response tendencies, which are how one reacts to the emotion in terms of your behaviors
(i.e., the tendency to feel happy, sad, etc.). Antecedent- and response-focused emotion regulation (e.g., suppression) modulate the final experiences of emotion to determine whether emotions are expressed or suppressed.

**Moving Forward**

Like many psychological measures, these four measures of emotional expressivity suffer from a major problem: reliance on averaged item answers to form one aggregate measure of the construct. Whereas averaging is convenient, such scaling techniques of classical test theory “do not guarantee that measurement precision will be equally distributed across the domain of interest” (Fraley et al., 2000, p. 350). That is, even though scores from each of the four scales demonstrate good if not excellent psychometric properties, such classical test theory properties (e.g., Cronbach’s alpha) leave more precision to be desired. For example, consider the 5-point Likert scale on the DDI ranging from 1 (*strongly disagree*) to 5 (*strongly agree*), with each point on the scale receiving a label. Fraley et al. assert that the differences between scale points of this Likert scale are ambiguous until analyses indicate otherwise. Specifically, the difference between endorsing a 1 (*strongly disagree*) and a 2 (*disagree*) does not necessarily represent the same decision in choosing between a 2 (*disagree*) and a 3 (*neither agree nor disagree*), or any other comparison of adjacent values. In classical test theory, researchers frequently ignore such scaling issues when determining test reliability and validity. In the next section I discuss item response theory as a remedy for this problem.
Item Response Theory

Item response theory (IRT) is a form of analysis that expounds on the concepts of classical test theory. In general, IRT analysis indicates if items and tests suffer from scaling problems. IRT analysis also addresses several problems of classical test theory.

Before discussing the functions and interpretations of IRT analysis, the reader must be aware of an assumption of this statistical procedure. A major assumption of IRT analysis is that people have a certain level of the latent trait, in this case emotional expressivity, measured by a test (Embretson & Reise, 2000). The latent trait is an unobservable variable that influences responses to items and, therefore, overall test scores. Researchers assume that the latent trait exists even though it is never directly measured. In addition, levels of the latent trait are both reflective and predictive of response patterns to items. Thus, IRT is also known as “latent trait theory” because the analysis focuses on understanding the assumed latent trait underlying responses at both the item level and the test level.

In graphical displays of IRT analyses, the latent trait is always the variable on the abscissa (see Figure 1). The notation for the latent trait is theta (θ), and the values of theta typically range from -3 to +3 in 1-unit increments, where 1 unit is 1 standard deviation (Fraley et al., 2000). Similar to the concept of z-scores, theta values indicate direction as well as magnitude. Direction refers to the sign of theta (i.e., positive or negative), whereas magnitude refers to the distance from theta = 0. A theta of 0 refers to individuals who score at the mean of the latent trait. Levels of theta less than 0 represent relative positions of individuals who score on the low end of the latent trait. Conversely,
theta levels greater than 0 represent relative positions of individuals who score on the high end of the latent trait.

Figure 1. Example of an item characteristic curve ranging from theta of -3 to theta of +3.

Several methods exist for conducting an item response analysis, the most common of which are versions of the logistic model. To analyze a self-report measure scored on a Likert scale, researchers often utilize the two-parameter logistic (2PL) model, the formula for which is:

$$P_j(\Theta_i) = \frac{1}{1 + \exp(-a_j(\Theta_i - b_j))}$$

(1)

where $P_j(\Theta_i)$ is the probability that an individual, $i$, with latent trait $\Theta_i$ will choose a response in the keyed direction on item $j$. The 2PL model estimates parameters of item difficulty ($b_j$) and item discrimination ($a_j$) (Baker, 2001). In typical item response analyses (i.e., an analysis of dichotomously scored items), *item difficulty* refers to the theta value at which the probability of answering an item correctly (or in the keyed direction) is 0.5. (The interpretation of item difficulty changes slightly for non-ability
tests such as personality measures.) *Item discrimination*, on the other hand, refers to how well an item differentiates people who are high on the latent trait from people low on the latent trait. IRT analysis under the 2PL model produces a series of graphs by which to interpret item and test properties, and these graphs make it easy to visualize item difficulty and item discrimination.

An item characteristic curve (ICC) is often the first graph generated in an IRT analysis. The ICC shows the relation between a person’s latent trait and the probability of item endorsement, specifically for higher response options (Fraley et al., 2000). ICCs are non-linear, monotonically increasing graphs that ideally form an S-shape, also called an ogive (see Figure 2). Item difficulty will always lie along the line where $p = 0.5$, whereas item discrimination will be reflected by the steepness of the curve. Figure 2 shows ideal ICCs for a set of simulated data scored on a 5-point Likert scale (e.g., 1 = *strongly disagree*, 5 = *strongly agree*). For a Likert scale (i.e., polytomously scored items), an ICC exists for each $(m - 1)$ dichotomy where $m$ is the number of response options for a scale. A 5-point Likert scale will therefore have $(5 - 1) = 4$ dichotomies. The first dichotomy compares a score of 1 to scores of 2 – 5. The second dichotomy compares scores of 1 and 2 to scores of 3 – 5. The third dichotomy compares scores of 1 – 3 to scores of 4 and 5. Finally, the fourth dichotomy compares scores of 1 – 4 to a score of 5. The interpretation of item difficulty therefore changes for polytomous response options: item difficulty is the theta value at which the probability of endorsing a response option compared to other response options is 0.5. For the first dichotomy, $p = 0.5$ aligns with theta = -1.5. This means that respondents with a latent trait of -1.5 will have a 0.5 probability of endorsing a 1 (*strongly disagree*) as compared to a 2 – 5.
The interpretation of item discrimination for polytomous response options still reflects the slope of the ICC. The ICCs in Figure 2 depict ideal item discrimination for each dichotomy. The general trend is for the probability of item endorsement to increase gradually, then quickly, and return to gradually increasing as theta increases from below average levels to above average levels. When each ICC has the same slope, item discrimination is the same for each dichotomy. The simulated graph displays an ideal situation in which the slopes are all the same. In typical data, each of the lines would instead have different slopes, and some slopes may overlap. Overlapping slopes mean that items have different item discrimination values. Extreme item discrimination can manifest as a step function, representing absolute discrimination at a certain theta value, or a near horizontal line, meaning theta values are of little or no use in understanding a person’s responses (Embretson & Reise, 2000). The ICCs relate to the next graph typically generated in an IRT analysis: an item information curve.

Item information curves (IICs) show the relation between a person’s latent trait and the probability of selecting a given response option on the scale (Fraley et al., 2000).
These graphs do not focus on dichotomies but rather provide a curve for each point on the measurement scale. Thus, a 5-point Likert scale would have five IICs shown in the graph. In addition, each IIC graph displays the relative information that each individual item provides, meaning that one graph exists for each item on the Likert scale. Figure 3 shows IICs for an item using simulated data. The ideal curvature of each line is unique. For a scale value of 1, defined as *strongly disagree*, people are very likely to endorse a 1 when they have very low levels of the latent trait. As a person’s latent trait increases to the mean, the probability of selecting a response option of 1 becomes much lower. When \( \theta = 1 \), the probability of selecting strongly disagree is almost 0. By following the lines for each response option, we gain an understanding of when people are more or less likely to endorse a given response option. For example, people with lower than average latent traits are most likely to endorse a 1 or 2, people with slightly below average levels of the latent trait are most likely to endorse a 3, and people with higher than average latent traits are most likely to endorse a 4 or 5 in the keyed direction. The intersections of the five IICs (i.e., \( \theta \) values of approximately -2, -1, 0, and 1), identified as thresholds, reflect the difficulty of the four dichotomies of the ICCs (Embretson & Reise, 2000).
Figure 3. Item information curves for each rating on a 5-point Likert scale.

The test information curve (TIC) is the sum of all IIC values across all scale items. The TIC graph shows the relation between a person’s latent trait and the overall precision of the scale, where precision means measurement precision across the latent trait. Figure 4 shows a TIC for a 10-item measure using simulated data. The ideal shape of this graph is depicted: test precision is low at very low (-3) and very high (+3) values of theta; test precision increases sharply from very low (-3) to moderately low (-1) and from very high (+3) to moderately high (+1) values of theta; and test precision plateaus for moderately low (-1) to moderately high (+1) values of theta. The amount of precision is therefore unique for different levels of the latent trait as well as dependent on the number of items in a test. This feature of IRT analysis builds on the indices of precision obtained from classical test theory such as reliability and validity. Oftentimes, internal consistency (i.e., Cronbach’s alpha) is the only measure of precision used to describe a measure in a research report. Researchers must therefore assume that scores from the measure are reliable for all participants. By constructing IICs and a TIC, researchers using IRT analysis overcome this assumption, and researchers can understand how
individual differences in the latent trait map onto the precision and overall understanding gained from item responses (Embretson & Reise, 2000).

![Figure 4. Test information curve for all 10 items on a 5-point Likert scale.](image)

**Purpose and Research Questions**

The purpose of the current study was to conduct an item response theory analysis of four emotional expressivity scales: the SCS, the DDI, the EES, and the ERQ-S. An example was provided by Fraley et al. (2000) in the analysis of attachment measures. Fraley et al. (2000) utilized IRT analysis to determine which measures of attachment provided good information across a range of latent trait levels. I followed an outline of their study, as well as the work of Toland (2014), who conducted an IRT analysis of a self-efficacy measure.

To complete this process, I first calculated the means and standard deviations of the measurement scales. I then conducted reliability analyses to obtain Cronbach’s alpha. The previous steps were necessary to replicate previous research as well as allow subsequent researchers to easily access all of my work. I then conducted four
confirmatory factor analyses to determine that each scale is unidimensional. Next, I examined parameters of item discrimination (i.e., the $a$ parameter) and item difficulty (i.e., the $b$ parameter) for the individual items. This process allowed me to see how informative the items were across the latent trait range and how likely items were to be endorsed by individuals with different latent trait levels. Finally, I examined the test information curves (TICs) for all test information across latent trait values. Specifically, I examined the TICs of the SCS, the DDI, the EES, and the ERQ-S. Examination of the TICs was necessary to ascertain whether or not the test adequately measured the latent trait.

For the SCS, previous research (Cramer & Barry, 1999, Study 2; Larson & Chastain, 1990) indicated that participants tend to score slightly lower than midpoint on the 5-point Likert scale. Therefore, I predicted the TIC would be positively skewed. For the DDI, previous research (Kahn & Hessling, 2001; Kahn et al., 2012) indicated that participants tend to score higher than midpoint on the 5-point Likert scale. Therefore, I predicted the TIC would be negatively skewed. For the EES, previous research (Kring et al., 1994) indicated that participants tend to score higher than midpoint on the 6-point Likert scale. Therefore, I predicted the TIC would be negatively skewed. Finally, the ERQ-S tended to produce scores lower than the midpoint (Gross & John, 2003; Kahn et al., 2012), so I predicted the TIC would be positively skewed.
CHAPTER III

METHOD

Participants

This study utilized archival data from two sources, both of which were collected from students attending a large Midwestern university. The first source (Kahn & Garrison, 2009) provided data from the DDI and ERQ-S and included 836 participants with a mean age of 19.49 (SD = 2.11). The majority of participants were female (77%), and ethnicity was represented as follows: Caucasian (85%), African-American (7%), Hispanic (3%), Asian-American (2%), and all others (3%). The second source (Barr et al., 2008) provided data from the SCS and the EES and included 552 participants with a mean age of 19.51 (SD = 1.85). The majority of participants were female (53%), and ethnicity was represented as follows: Caucasian (88%), African-American (5%), Hispanic (2%), Asian-American (2%), and all others (3%).

Measures

These instruments are described in detail in Chapter 2.

Self-Concealment

The 10 items of the Self-Concealment Scale measure active concealment, or hiding information, from others. Scores were measured on a 1 (strongly disagree) to 5 (strongly agree) Likert scale with low scores indicating low tendency to conceal and high scores indicating high tendency to conceal (Larson & Chastain, 1990).
Distress Disclosure

The Distress Disclosure Index (DDI) is a 12-item bipolar scale that measures degree of disclosure or concealment of personally distressing information. Scores were measured on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree). Low scores indicated rare disclosure of distressing information, and high scores indicated frequent disclosure of distressing information (Kahn & Hessling, 2001).

Emotional Expressivity

The Emotional Expressivity Scale (EES) is a 17-item scale that measures general emotional expressivity on a Likert scale from 1 (never true) to 6 (always true). Appropriate items were reverse-coded and added together such that low scores indicated low emotional expressivity and high scores indicated high emotional expressivity (Kring et al., 1994).

Suppression

The Emotional Regulation Questionnaire – Suppression subscale measures behavioral inhibition from expressing positive and negative emotions. This 4-item scale is measured on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree), with low scores indicating low suppression and high scores indicating high suppression (Gross & John, 2003).

Procedure

For the first archival source (Kahn & Garrison, 2009), participants completed the survey in the lab in testing groups of up to 30 people. Participants provided informed consent before completing the DDI and ERQ-S along with other questionnaires not
pertinent to the current study. Once finished, the participants were read a debriefing statement and received extra credit in a psychology course.

For the second archival source (Barr et al., 2008), participants completed the survey in a group testing session. Participants provided informed consent before completing the SCS and the EES along with other questionnaires not pertinent to the current study. After completion of the questionnaires, the participants were read a debriefing statement and received extra credit in a psychology course.

**Data Analysis**

This study used an item response theory (IRT) analysis. I deleted participants who did not complete the entire survey. The final sample for each scale was as follows: 551 participants fully completed the SCS, 829 participants fully completed the DDI, 547 participants fully completed the EES, and 824 participants fully completed the ERQ-S. I then followed steps outlined in Toland (2014) to analyze the data. First, I clarified the purpose of the study (as outlined in Chapter Two). Next, I considered the relevant models and chose the graded response model. Subsequently, I used several statistical packages to conduct preliminary data inspection. I computed the means and standard deviations for each of the four scales on which data were collected, which was necessary for comparison to previous research. I conducted reliability analyses for each of the four scales, which was again necessary for comparison to previous research. I then used LISREL (Jöreskog & Sörbom, 1999) to conduct a confirmatory factor analysis for each of the four scales. Determining the dimensionality of each scale was necessary because an assumption of this IRT analysis is unidimensionality of the scale. In the case that a
scale did not meet the assumption of unidimensionality, it was excluded from further analyses.

Finally, I utilized IRTPRO (*Item Response Theory for Patient-Reported Outcomes*) version 3.1 (Cai, du Toit, & Thissen, 2015) to conduct the IRT analysis. An IRT analysis included the following: determining the number of responses for each response category for each item for each scale, calculating item difficulty and item discrimination parameters for each item for each scale, computing item information for each item for each scale, and computing test information for each scale. I generated graphs of the item and test information using Toland’s (2014) supplemental materials.
CHAPTER IV
RESULTS

Descriptive Statistics

Many steps are involved in completing an IRT analysis. First, calculation of the means and standard deviations for each scale are necessary to compare to previous research. I determined that the means and standard deviations for each of the four scales were somewhat comparable to the previous literature (see Table 1). That is, the SCS was similar to that of previous research because previous research ranged from a low mean of 25.92 and low standard deviation of 7.30 (Larson & Chastain, 1990) to a high mean of 30.16 and high standard deviation of 9.19 (Cramer & Barry, 1999, Study 2). The DDI was slightly higher and slightly more variable than that of previous research, in which the mean and standard deviation ranged from 39.93 and 9.51 (Kahn & Hessling, 2001) to a mean and standard deviation of 42.16 and 9.55 (Kahn et al., 2012), respectively. The EES was slightly higher and slightly more variable compared to previous literature, which ranged from a low mean of 61.18 and low standard deviation of 12.04 (Kring et al., 1994, Study 3) to a high mean of 64.67 and a high standard deviation of 13.59 (Kring et al., 1994, Study 1; Kring et al., 1994, Study 4). Finally, the ERQ-S was slightly lower and slightly more variable compared to previous literature. In previous literature, means and standard deviations ranged from 12.77 and 4.35 (Kahn et al., 2012) to 13.56 and 4.58...
(Gross & John, 2003), respectively. Despite small aberrations from previous literature, it is acceptable to proceed with the analyses.

Table 1

Means, Standard Deviations, and Reliabilities for Study Scales

<table>
<thead>
<tr>
<th>Scale</th>
<th>Final N</th>
<th>M</th>
<th>SD</th>
<th>α</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-Concealment Scale</td>
<td>551</td>
<td>26.50</td>
<td>9.09</td>
<td>0.86</td>
</tr>
<tr>
<td>Distress Disclosure Index</td>
<td>829</td>
<td>42.40</td>
<td>11.66</td>
<td>0.95</td>
</tr>
<tr>
<td>Emotional Expressivity Scale</td>
<td>547</td>
<td>67.35</td>
<td>15.31</td>
<td>0.92</td>
</tr>
<tr>
<td>Emotional Regulation Questionnaire –</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Suppression Subscale</td>
<td>824</td>
<td>11.75</td>
<td>4.95</td>
<td>0.77</td>
</tr>
</tbody>
</table>

**Reliability Coefficients**

Next, computation of reliability estimates is necessary for use in subsequent research. I found reliability estimates, as computed by Cronbach’s alpha, to be similar to previous research (see Table 1). That is, scores from the SCS fell within the expected range of .83 – .87 (Larson & Chastain, 1990). The DDI scores also fell within the expected range of .92 – .85 (Kahn & Hessling, 2001). Next, the EES scores fell within the expected range of .90 – .93 (Kring et al., 1994). Lastly, the ERQ-S scores were slightly more reliable than the expected value of .73 (Gross & John, 2003). Overall, reliability estimates were strong.
CFAs for the SCS, DDI, EES, and ERQ-S

Unidimensionality is an assumption of IRT. Therefore, CFAs for each scale were necessary to determine the dimensionality of the scale. Using LISREL, I conducted four CFAs: one for the SCS, one for the DDI, one for the EES, and one for the ERQ-S. The following fit indices and cut-offs for the fit indices were utilized (Hu and Bentler, 1999): the Satorra-Bentler $\chi^2$, the root mean square error of approximation (RMSEA) indicated a good fit if it was less than .06; the Comparative Fit Index (CFI) indicated a good fit if it was greater than .95; the Non-Normed Fit Index (NNFI) indicated a good fit if it was greater than .95; and the standardized root mean square residual (SRMR) indicated a good fit if it was less than .08.

For the SCS, I specified a model with a single latent factor and all 10 items loading on the factor. All factor loadings were significant, and standardized factor loadings ranged from .54 to .78. The fit statistics indicated that the model provided a good fit to the data based on the CFI, NNFI, and SRMR (see Table 2). The RMSEA was marginally acceptable. Given the other fit statistics, however, I concluded that there is no reason to think the SCS is not unidimensional.
<table>
<thead>
<tr>
<th>Scale</th>
<th>S-B $\chi^2$</th>
<th>df</th>
<th>RMSEA</th>
<th>90% CI</th>
<th>CFI</th>
<th>NNFI</th>
<th>SRMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCS</td>
<td>155.12</td>
<td>35</td>
<td>0.08</td>
<td>[.07, .09]</td>
<td>0.98</td>
<td>0.97</td>
<td>0.05</td>
</tr>
<tr>
<td>DDI</td>
<td>342.66</td>
<td>54</td>
<td>0.08</td>
<td>[.07, .09]</td>
<td>0.99</td>
<td>0.99</td>
<td>0.04</td>
</tr>
<tr>
<td>EES</td>
<td>990.51</td>
<td>119</td>
<td>0.09</td>
<td>[.09, .10]</td>
<td>0.96</td>
<td>0.96</td>
<td>0.07</td>
</tr>
<tr>
<td>ERQ-S</td>
<td>2.33</td>
<td>2</td>
<td>0.01</td>
<td>[.00, .07]</td>
<td>1.00</td>
<td>1.00</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Note. S-B $\chi^2$ = Satorra-Bentler Scaled chi-square, df = degrees of freedom associated with Satorra-Bentler chi-square, RMSEA = root mean square error of approximation, CFI = comparative fit index, NNFI = non-normed fit index, SRMR = standardized root mean square residual.

For the DDI, I specified a model with a single latent factor and all 12 items loading on the factor. All factor loadings were significant, and standardized factor loadings ranged from .75 to .88. The fit statistics indicated that the model provided a good fit to the data based on the CFI, NNFI, and SRMR (see Table 2). The RMSEA was marginally acceptable, but given the other fit statistics I concluded that the DDI items are consistent with the assumption of unidimensionality.

For the EES, I specified a model with a single latent factor and all 17 items loading on the factor. All factor loadings were significant, and standardized factor loadings ranged from .43 to .79. The low standardized factor loading could be a result of an overall bad item. The fit statistics indicated that the model provided a good fit to the
data based on the CFI and the NNFI (see Table 2). The SRMR and the RMSEA were marginally acceptable. Overall, I decided to proceed with caution that the EES items are consistent with the assumption of unidimensionality.

For the ERQ-S, I specified a model with a single latent factor and all 4 items loading on the factor. All factor loadings were significant, and the standardized factor loadings ranged from .59 to .90. The fit statistics indicated that the model provided excellent fit to the data based on the RMSEA, CFI, NNFI, and the SRMR (see Table 2). It should be noted that the degrees of freedom for the four scales’ models ranged from 2 (ERQ-S) to 119 (EES). The higher the degrees of freedom, the greater the parsimony of the model. Some fit statistics favor parsimony whereas some fit statistics ignore parsimony, such that the higher values of the RMSEA and SRMR are not unexpected for the EES given 119 degrees of freedom. Therefore, I concluded that all four measures of emotional expressivity met the assumption for unidimensionality, and I proceeded with analyses.

**IRT Analysis for the SCS**

The first instrument analyzed with IRT analyses was the SCS. Scores from the SCS yielded means and standard deviations that were similar to those in past research, the reliability of SCS scores fell within an acceptable range, and the CFA did not show any violations to the assumption of unidimensionality. The first step in the IRT analyses was data inspection to determine that a minimum of 5% (Toland, 2014) of responses fell within each of the response categories. Because the minimum percentage of responses that fell within a response option for the SCS was 6%, this justified the use of all five response categories.
Next, item calibrations were calculated using IRTPRO. See Table 3 for all item calibration results for the graded response model fit to the 10-item five-category SCS. Recall that the $a$ parameter is item discrimination. Item discrimination parameters ranged from 1.13 (Item 7) to 2.45 (Item 4). Variation in item discrimination parameters suggested that unique item discrimination parameters from the 2PL model are appropriate for these data. Item difficulty parameters (the $b$ columns in Table 3) ranged from -1.51 to -0.27 for $b_1$, -0.23 to 0.54 for $b_2$, 0.32 to 1.39 for $b_3$, and 1.03 to 2.18 for $b_4$. 
Table 3

Graded Response Model Item Parameter Estimates and Item-Fit Statistics for 10-Item Five-Category Self-Concealment Scale

<table>
<thead>
<tr>
<th>Item</th>
<th>(a)</th>
<th>(b_1)</th>
<th>(b_2)</th>
<th>(b_3)</th>
<th>(b_4)</th>
<th>(S-\chi^2)</th>
<th>(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.58 (0.14)</td>
<td>-0.68 (0.10)</td>
<td>-0.06 (0.08)</td>
<td>0.32 (0.07)</td>
<td>1.11 (0.10)</td>
<td>103.33</td>
<td>0.3108</td>
</tr>
<tr>
<td>2</td>
<td>1.52 (0.14)</td>
<td>-0.38 (0.09)</td>
<td>0.54 (0.08)</td>
<td>1.39 (0.12)</td>
<td>2.15 (0.17)</td>
<td>104.64</td>
<td>0.3041</td>
</tr>
<tr>
<td>3</td>
<td>1.62 (0.13)</td>
<td>-1.51 (0.13)</td>
<td>-0.23 (0.08)</td>
<td>0.62 (0.08)</td>
<td>1.53 (0.12)</td>
<td>89.60</td>
<td>0.6910</td>
</tr>
<tr>
<td>4</td>
<td>2.45 (0.21)</td>
<td>-0.27 (0.07)</td>
<td>0.39 (0.06)</td>
<td>0.82 (0.07)</td>
<td>1.43 (0.09)</td>
<td>75.45</td>
<td>0.7102</td>
</tr>
<tr>
<td>5</td>
<td>1.56 (0.13)</td>
<td>-1.14 (0.12)</td>
<td>-0.03 (0.08)</td>
<td>0.95 (0.09)</td>
<td>2.18 (0.17)</td>
<td>144.59</td>
<td>0.0005</td>
</tr>
<tr>
<td>6</td>
<td>1.94 (0.16)</td>
<td>-0.59 (0.09)</td>
<td>0.31 (0.07)</td>
<td>0.91 (0.08)</td>
<td>1.74 (0.12)</td>
<td>106.67</td>
<td>0.1107</td>
</tr>
<tr>
<td>7</td>
<td>1.13 (0.11)</td>
<td>-1.21 (0.14)</td>
<td>-0.09 (0.09)</td>
<td>0.85 (0.11)</td>
<td>2.03 (0.19)</td>
<td>124.07</td>
<td>0.1867</td>
</tr>
<tr>
<td>8</td>
<td>2.20 (0.19)</td>
<td>-0.37 (0.07)</td>
<td>0.15 (0.06)</td>
<td>0.51 (0.06)</td>
<td>1.03 (0.08)</td>
<td>107.00</td>
<td>0.0621</td>
</tr>
<tr>
<td>9</td>
<td>2.11 (0.18)</td>
<td>-0.40 (0.08)</td>
<td>0.43 (0.06)</td>
<td>1.08 (0.08)</td>
<td>1.96 (0.13)</td>
<td>100.16</td>
<td>0.1964</td>
</tr>
<tr>
<td>10</td>
<td>1.24 (0.12)</td>
<td>-0.70 (0.11)</td>
<td>0.20 (0.09)</td>
<td>0.93 (0.10)</td>
<td>2.05 (0.18)</td>
<td>99.55</td>
<td>0.6832</td>
</tr>
</tbody>
</table>

*Note. \(a\) = item discrimination parameter, \(b\) = item difficulty parameter, \(S-\chi^2\) = item-fit statistic. Values in parenthesis are item parameter standard error estimate.
The next step was to test the assumption of local independence by investigating the standardized local dependency $\chi^2$ values for each of the 45 item pairs (i.e., Item 1 with Item 2, Item 1 with Item 3, etc.). Standardized local dependency $\chi^2$ values discern whether a response to an item is independent from a response to any other item. I used the suggested value of |10| (Chen & Thissen, 1997) to determine whether the local dependency statistics were large and likely to represent issues with local dependence. All but one of the values (i.e., the intersection of Item 6 and Item 9) were within the suggested range which suggested no significant violation of local independence of the data.

I then tested the functional form by visually inspecting the 10 plots of the trace line graphs (i.e., the IICs) generated in IRTPRO. Overall, the plots (which, for brevity, are included in the Appendix) followed the ideal form identified in Chapter 2 (i.e., lower response options were mostly endorsed by people with lower than average latent traits, etc.) with a few exceptions. Item 1 appeared not to benefit from the inclusion of the middle response option. On the other hand, Item 2 might benefit from the combination of the two highest response options. Items 4, 6, 7, 8, and 9 also appeared not to benefit from the inclusion of the middle response option. Overall, the IICs suggest the middle response option (i.e., “do not disagree or agree”) may not be needed for the SCS.

Inspection of the model-data fit for items and the overall model occurred next. Model-data fit refers to statistical comparisons between the observed data and data the IRT model would predict. For items, I inspected the S-$\chi^2$ values presented in Table 3. A statistically significant value means that the model does not fit a given item. Although it appears that the model is not a good fit to Item 5, the fact that only 1 of the 10 scale items
was not a good fit is still satisfactory. In other words, 9 of the 10 scale items were well represented by the estimated item parameters for the graded response model. For the overall graded response model level fit, there was adequate model-data fit given the RMSEA of 0.04. It therefore appears that the model fits both the items and the SCS as a whole.

Finally, I examined the TIC for the SCS (see Figure 5). The TIC revealed that the SCS provides good information for theta values of approximately -1 to 2. This means the scale does not capture the low end of self-concealment as well as it captures the high end of self-concealment. There is also a very slight negative skew in the truncated range of the TIC. Caution should be used when interpreting test information outside of the range of -1 to 2.

![Test information curve for 10-item Self-Concealment Scale.](image)

*Figure 5. Test information curve for 10-item Self-Concealment Scale.*
IRT Analysis for the DDI

The second instrument analyzed with IRT analyses was the DDI. Similar to the SCS, scores from the DDI yielded consistent means and standard deviations to those of previous research; the reliability of DDI scores were acceptable; and the CFA did not show any violations to the assumption of unidimensionality. First, I inspected the data to determine that a minimum of 5% (Toland, 2014) of responses fell within each of the response categories. Unlike the SCS, 4 of the DDI items did not have a minimum of 5% of responses within each response category. For this reason, I deemed it necessary to combine the two lowest response categories to meet this recommendation. That is, responses of 1 (“strongly disagree”) and 2 (“disagree”) were combined to both reflect a 2. Subsequent analyses of the DDI will reflect a scale that ranges from 2 to 5.

Next, I calculated item calibrations for the graded response model fit to the 12-item recoded four-category DDI (see Table 4). Item discrimination parameters ranged from 2.17 (Item 7) to 3.85 (Item 9). Such large item discrimination parameters suggest the items discriminate very well (Baker, 2001). Item difficulty parameters ranged from -1.31 to -0.37 for $b_1$, -0.66 to 0.10 for $b_2$, and 0.25 to 1.05 for $b_3$.  

41
Table 4

Graded Response Model Item Parameter Estimates and Item-Fit Statistics for 12-Item Recoded Four-Category Distress Disclosure Index

<table>
<thead>
<tr>
<th>Item</th>
<th>$a$</th>
<th>$b1$</th>
<th>$b2$</th>
<th>$b3$</th>
<th>S-$\chi^2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.24 (0.14)</td>
<td>-1.31 (0.08)</td>
<td>-0.66 (0.06)</td>
<td>0.25 (0.05)</td>
<td>87.37</td>
<td>0.2428</td>
</tr>
<tr>
<td>2</td>
<td>2.72 (0.16)</td>
<td>-0.69 (0.06)</td>
<td>-0.17 (0.05)</td>
<td>0.82 (0.06)</td>
<td>105.38</td>
<td>0.0063</td>
</tr>
<tr>
<td>3</td>
<td>2.99 (0.18)</td>
<td>-1.15 (0.07)</td>
<td>-0.51 (0.05)</td>
<td>0.43 (0.05)</td>
<td>59.77</td>
<td>0.7229</td>
</tr>
<tr>
<td>4</td>
<td>3.32 (0.20)</td>
<td>-0.82 (0.06)</td>
<td>-0.31 (0.05)</td>
<td>0.58 (0.05)</td>
<td>90.88</td>
<td>0.0277</td>
</tr>
<tr>
<td>5</td>
<td>2.95 (0.18)</td>
<td>-0.37 (0.05)</td>
<td>0.02 (0.04)</td>
<td>0.87 (0.06)</td>
<td>64.49</td>
<td>0.5303</td>
</tr>
<tr>
<td>6</td>
<td>3.26 (0.20)</td>
<td>-0.69 (0.05)</td>
<td>-0.06 (0.04)</td>
<td>0.81 (0.05)</td>
<td>104.04</td>
<td>0.0015</td>
</tr>
<tr>
<td>7</td>
<td>2.17 (0.13)</td>
<td>-0.79 (0.06)</td>
<td>-0.12 (0.05)</td>
<td>0.95 (0.06)</td>
<td>94.12</td>
<td>0.1336</td>
</tr>
<tr>
<td>8</td>
<td>2.36 (0.14)</td>
<td>-0.99 (0.07)</td>
<td>-0.28 (0.05)</td>
<td>0.86 (0.06)</td>
<td>64.43</td>
<td>0.8255</td>
</tr>
<tr>
<td>9</td>
<td>3.85 (0.24)</td>
<td>-0.87 (0.06)</td>
<td>-0.33 (0.04)</td>
<td>0.58 (0.05)</td>
<td>96.70</td>
<td>0.0011</td>
</tr>
<tr>
<td>10</td>
<td>2.73 (0.16)</td>
<td>-1.06 (0.07)</td>
<td>-0.36 (0.05)</td>
<td>0.67 (0.05)</td>
<td>88.12</td>
<td>0.0952</td>
</tr>
<tr>
<td>11</td>
<td>2.85 (0.17)</td>
<td>-0.60 (0.05)</td>
<td>0.10 (0.05)</td>
<td>1.05 (0.06)</td>
<td>99.54</td>
<td>0.0060</td>
</tr>
<tr>
<td>12</td>
<td>2.45 (0.14)</td>
<td>-1.00 (0.07)</td>
<td>-0.20 (0.05)</td>
<td>0.99 (0.06)</td>
<td>86.63</td>
<td>0.1492</td>
</tr>
</tbody>
</table>

Note. $a$ = item discrimination parameter, $b$ = item difficulty parameter, S-$\chi^2$ = item-fit statistic. Values in parenthesis are item parameter standard error estimate.

To test the assumption of local independence, I investigated the standardized local dependency $\chi^2$ values for each of the 66 item pairs of the DDI. Several of the local dependency $\chi^2$ values (i.e., 18) were not within the suggested range, and the highest value of 24.3 occurred at the crossing of Item 4 with Item 9. Such a large number of values
above the recommended threshold suggests there may be violations to the assumption of local independence of the data.

I then visually inspected each of the 12 item’s IICs to assess the functional form of the DDI. The IICs revealed several issues with the items. The response option of 3 (“neither agree nor disagree”) does not appear to be beneficial for Items 1, 2, 4, 5, and 7. It does appear, however, that combining the two lowest response options was beneficial in the construction of the IICs. After combining the two lowest response options, it still seems that the DDI does not significantly benefit from the inclusion of a neutral response option.

Next, I examined the item and model-data fit by reviewing the S-χ² values that are presented in Table 4. It appears that 5 of the items (i.e., Items 2, 4, 6, 9, and 11) are a poor fit to the model because they are significant. Having 5 out of 12 items with significant S-χ² values may suggest the items are not well represented by the estimated item parameters for the graded response model. In terms of the overall graded response model, the low RMSEA value of 0.05 indicates good model-data fit. Therefore, item and model-data fit results for the DDI are inconsistent.

Finally, examination of the TIC for the DDI was necessary (see Figure 6). The TIC revealed that the DDI provides good information for theta values of approximately -1.5 to 1.5, meaning the scale captures moderately low and moderately high levels of distress disclosure. It is best to interpret test information within this range.
The next instrument analyzed with IRT analyses was the EES. Again, the necessary steps were conducted prior to the IRT analysis. I then followed the same steps outlined in the prior analyses. First, I inspected the data to determine that a minimum of 5% (Toland, 2014) of responses fell within each of the response categories. Because most items (13 out of 17) did not reach a minimum of 5% of responses for the first response category, the use of all six response categories was not justified. Instead, response categories of 1 and 2 were combined to reflect a response option of 2. For all subsequent analyses, the EES scale now ranges from 2 to 6, where a 2 encompasses both “never true of me” and “rarely true of me.”

I then calculated item calibrations for the graded response model fit to the 17-item five-category recoded EES (see Table 5). Item discrimination parameters ranged from...
0.87 (Item 4) to 2.48 (Item 15). Variation in item discrimination parameters suggested that unique item discrimination parameters from the 2PL model may be appropriate for these data. Item difficulty parameters ranged from -2.91 to -0.49 for $b_1$, -1.44 to 0.25 for $b_2$, -0.48 to 1.15 for $b_3$, and 0.80 to 2.46 for $b_4$. 
Table 5

Graded Response Model Item Parameter Estimates and Item-Fit Statistics for 17-Item Recoded Five-Category Emotional Expressivity Scale

<table>
<thead>
<tr>
<th>Item</th>
<th>$a$ (SD)</th>
<th>$b1$ (SD)</th>
<th>$b2$ (SD)</th>
<th>$b3$ (SD)</th>
<th>$b4$ (SD)</th>
<th>S-χ² (SD)</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.02 (0.25)</td>
<td>-1.37 (0.19)</td>
<td>-0.54 (0.11)</td>
<td>0.28 (0.06)</td>
<td>1.49 (0.15)</td>
<td>122.26 (0.15)</td>
<td>0.4770</td>
</tr>
<tr>
<td>2</td>
<td>1.63 (0.24)</td>
<td>-1.97 (0.30)</td>
<td>-1.18 (0.19)</td>
<td>-0.48 (0.10)</td>
<td>0.80 (0.12)</td>
<td>224.22 (0.12)</td>
<td>0.0001</td>
</tr>
<tr>
<td>3</td>
<td>1.80 (0.25)</td>
<td>-1.20 (0.19)</td>
<td>-0.39 (0.09)</td>
<td>0.38 (0.07)</td>
<td>1.68 (0.21)</td>
<td>138.55 (0.21)</td>
<td>0.1920</td>
</tr>
<tr>
<td>4</td>
<td>0.87 (0.13)</td>
<td>-2.91 (0.42)</td>
<td>-1.44 (0.22)</td>
<td>0.10 (0.10)</td>
<td>1.94 (0.28)</td>
<td>148.73 (0.28)</td>
<td>0.7470</td>
</tr>
<tr>
<td>5</td>
<td>1.65 (0.24)</td>
<td>-1.26 (0.20)</td>
<td>-0.43 (0.10)</td>
<td>0.60 (0.09)</td>
<td>1.93 (0.25)</td>
<td>152.15 (0.25)</td>
<td>0.0894</td>
</tr>
<tr>
<td>6</td>
<td>2.39 (0.32)</td>
<td>-1.24 (0.20)</td>
<td>-0.44 (0.10)</td>
<td>0.45 (0.06)</td>
<td>1.55 (0.17)</td>
<td>102.10 (0.17)</td>
<td>0.7384</td>
</tr>
<tr>
<td>7</td>
<td>1.70 (0.23)</td>
<td>-1.45 (0.24)</td>
<td>-0.57 (0.13)</td>
<td>0.33 (0.06)</td>
<td>1.66 (0.19)</td>
<td>147.93 (0.19)</td>
<td>0.1937</td>
</tr>
<tr>
<td>8</td>
<td>1.32 (0.16)</td>
<td>-0.49 (0.11)</td>
<td>0.04 (0.08)</td>
<td>0.75 (0.10)</td>
<td>1.77 (0.20)</td>
<td>135.21 (0.20)</td>
<td>0.6446</td>
</tr>
<tr>
<td>9</td>
<td>2.02 (0.26)</td>
<td>-1.57 (0.24)</td>
<td>-0.72 (0.14)</td>
<td>0.07 (0.06)</td>
<td>1.31 (0.13)</td>
<td>141.61 (0.13)</td>
<td>0.0771</td>
</tr>
<tr>
<td>10</td>
<td>2.13 (0.27)</td>
<td>-1.49 (0.23)</td>
<td>-0.78 (0.15)</td>
<td>0.15 (0.06)</td>
<td>1.42 (0.13)</td>
<td>120.81 (0.13)</td>
<td>0.3130</td>
</tr>
</tbody>
</table>

(Table continues)
<table>
<thead>
<tr>
<th>Item</th>
<th>$a$</th>
<th>$b1$</th>
<th>$b2$</th>
<th>$b3$</th>
<th>$b4$</th>
<th>$S-\chi^2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>2.27 (0.26)</td>
<td>-1.38 (0.22)</td>
<td>-0.65 (0.13)</td>
<td>-0.11 (0.08)</td>
<td>1.06 (0.09)</td>
<td>149.64</td>
<td>0.0193</td>
</tr>
<tr>
<td>12</td>
<td>2.40 (0.26)</td>
<td>-1.49 (0.22)</td>
<td>-0.75 (0.14)</td>
<td>0.01 (0.07)</td>
<td>1.14 (0.09)</td>
<td>118.87</td>
<td>0.3830</td>
</tr>
<tr>
<td>13</td>
<td>1.36 (0.18)</td>
<td>-0.59 (0.13)</td>
<td>0.20 (0.07)</td>
<td>1.12 (0.13)</td>
<td>2.46 (0.27)</td>
<td>158.16</td>
<td>0.1532</td>
</tr>
<tr>
<td>14</td>
<td>1.52 (0.20)</td>
<td>-0.56 (0.13)</td>
<td>0.25 (0.07)</td>
<td>1.15 (0.12)</td>
<td>2.19 (0.23)</td>
<td>115.33</td>
<td>0.8001</td>
</tr>
<tr>
<td>15</td>
<td>2.48 (0.25)</td>
<td>-1.42 (0.22)</td>
<td>-0.73 (0.14)</td>
<td>0.05 (0.07)</td>
<td>1.12 (0.08)</td>
<td>130.96</td>
<td>0.0657</td>
</tr>
<tr>
<td>16</td>
<td>1.27 (0.14)</td>
<td>-1.88 (0.28)</td>
<td>-0.65 (0.15)</td>
<td>0.71 (0.09)</td>
<td>2.22 (0.21)</td>
<td>158.77</td>
<td>0.1889</td>
</tr>
<tr>
<td>17</td>
<td>2.25 (0.25)</td>
<td>-1.11 (0.19)</td>
<td>-0.38 (0.11)</td>
<td>0.33 (0.06)</td>
<td>1.30 (0.09)</td>
<td>120.10</td>
<td>0.4030</td>
</tr>
</tbody>
</table>

*Note. $a$ = item discrimination parameter, $b$ = item difficulty parameter, $S-\chi^2$ = item-fit statistic. Values in parenthesis are item*
I next tested the assumption of local independence. The highest problematic value was the intersection of Item 7 and Item 11 (24.6). In total, there were 10 problematic values, suggesting violations of local independence of the data.

Then I tested the functional form of the IICs. Visual inspection of the 17 plots of the trace line graphs revealed several issues with the IICs. Items 2, 8, and 13 do not appear to benefit from the inclusion of response options of 3 (“occasionally true of me”) or 4 (“often true of me”). Similarly, a response option of 3 does not appear to benefit Items 5, 10, and 14. Overall, several of the IICs suggest the middle response options may not be necessary for the EES.

Following an assessment of functional form, I examined the model-data fit for items and the overall model. The S-\( \chi^2 \) values are presented in Table 5. It appears that Items 2 and 11 are not a good fit to the model because they are significant. Although it appears that the model is not a good fit to Items 2 and 11, the fact that only 2 of the 17 scale items were not a good fit is still satisfactory. In other words, 15 of the 17 scale items were well represented by the estimated item parameters for the graded response model. For the overall graded response model, the RMSEA of 0.11 indicates there was not adequate model-data fit. Thus, model-data fit results for the EES are mixed.

Lastly, I examined the TIC for the EES (see Figure 7). The TIC revealed that the EES is slightly positively skewed and provides good information for theta values of approximately -2 to 2, which means the scale captures both the low and the high end of emotional expressivity quite well. Interpreting information outside of the -2 to 2 range of theta is advisable with caution.
The final instrument analyzed with IRT analyses was the ERQ-S. Data inspection revealed that 2 of the 4 items did not meet the recommended minimum of 5% (Toland, 2014) of responses within each of the response categories. More specifically, a response option of 7 (“strongly agree”) did not reach 5% of responses for Items 2 and 3, and the use of all seven response categories was not justified. Instead, I combined response categories of 6 and 7 to reflect a response option of 6. The ERQ-S scale therefore ranged from 1 to 6 for all subsequent analyses. Even after this combination, however, Item 2 did not meet the recommended 5%. I justified the decision to keep a recoded 1 to 6 scale so as not to lose valuable information from the other items in which all other response options were selected a minimum amount of the time. Therefore, I proceeded with caution in interpreting the following results.

Figure 7. Test information curve for 17-item Emotional Expressivity Scale.

IRT Analysis for the ERQ-S

The final instrument analyzed with IRT analyses was the ERQ-S. Data inspection revealed that 2 of the 4 items did not meet the recommended minimum of 5% (Toland, 2014) of responses within each of the response categories. More specifically, a response option of 7 (“strongly agree”) did not reach 5% of responses for Items 2 and 3, and the use of all seven response categories was not justified. Instead, I combined response categories of 6 and 7 to reflect a response option of 6. The ERQ-S scale therefore ranged from 1 to 6 for all subsequent analyses. Even after this combination, however, Item 2 did not meet the recommended 5%. I justified the decision to keep a recoded 1 to 6 scale so as not to lose valuable information from the other items in which all other response options were selected a minimum amount of the time. Therefore, I proceeded with caution in interpreting the following results.
The next step involved calculating item calibrations for the graded response
model fit to the 4-item six-category recoded ERQ-S. See Table 6 for item discrimination
parameters, which ranged from 1.29 (Item 4) to 3.72 (Item 6). Variation in item
discrimination parameters suggested that unique item discrimination parameters from the
2PL model may be appropriate for these data. Item difficulty parameters are also
presented in Table 6, and they ranged from -1.89 to -0.49 for $b_1$, -0.79 to 1.01 for $b_2$, 0.18
to 1.94 for $b_3$, 0.76 to 2.71 for $b_4$, and 1.28 to 3.52 for $b_5$. 
Table 6

Graded Response Model Item Parameter Estimates and Item-Fit Statistics for 4-Item Recoded Six-Category Emotion Regulation Questionnaire – Suppression Subscale

<table>
<thead>
<tr>
<th>Item</th>
<th>$a$</th>
<th>$b_1$</th>
<th>$b_2$</th>
<th>$b_3$</th>
<th>$b_4$</th>
<th>$b_5$</th>
<th>$S-\chi^2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.64 (0.22)</td>
<td>-1.18 (0.07)</td>
<td>-0.25 (0.05)</td>
<td>0.31 (0.05)</td>
<td>0.76 (0.06)</td>
<td>1.28 (0.07)</td>
<td>50.06</td>
<td>0.5514</td>
</tr>
<tr>
<td>2</td>
<td>1.29 (0.10)</td>
<td>-0.49 (0.07)</td>
<td>1.01 (0.09)</td>
<td>1.94 (0.14)</td>
<td>2.71 (0.20)</td>
<td>3.52 (0.29)</td>
<td>102.77</td>
<td>0.0015</td>
</tr>
<tr>
<td>3</td>
<td>3.72 (0.45)</td>
<td>-0.68 (0.05)</td>
<td>0.04 (0.04)</td>
<td>0.54 (0.05)</td>
<td>1.00 (0.06)</td>
<td>1.51 (0.08)</td>
<td>70.90</td>
<td>0.0174</td>
</tr>
<tr>
<td>4</td>
<td>1.34 (0.10)</td>
<td>-1.89 (0.13)</td>
<td>-0.79 (0.08)</td>
<td>0.18 (0.06)</td>
<td>1.00 (0.09)</td>
<td>1.79 (0.13)</td>
<td>121.59</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

Note. $a$ = item discrimination parameter, $b$ = item difficulty parameter, $S-\chi^2$ = item-fit statistic. Values in parenthesis are item parameter standard error estimate.
I then tested the local independence. I determined that none of the LD statistics represent issues with local dependence.

Next, I tested the functional form of the IICs. Visual inspection of the 4 plots suggested issues with the IICs. For one, Item 1 does not appear to benefit much from recoded response options of 3 or 4. Item 2 is similar in that it does not benefit from the inclusion of recoded response options of 3, 4, or 5 in the given range of theta. Finally, the recoded response option of 4 is of no use for Item 4. Therefore, there are several issues with the recoded ERQ-S and the use of several response options.

Subsequently, I examined the model-data fit for items and the overall model (see Table 6 for S-$\chi^2$ values). Three of the 4 items are a bad fit to the model. That is, the S-$\chi^2$ values for Items 2, 3, and 4 were significant. Significant values suggest the 4 scale items were not well represented by the estimated item parameters for the graded response model. For the overall graded response model, the RMSEA of 0.04 indicates there was acceptable model-data fit. The model-data fit results for the ERQ-S are therefore mixed.

I finally examined the TIC for the ERQ-S (see Figure 8). The TIC suggests that the ERQ-S is slightly negatively skewed and provides good information for theta values of approximately -1.5 to 2, meaning the scale captures moderately low to high instances of suppression. The information outside of this range is not quite so helpful.
Figure 8. Test information curve for 4-item Emotion Regulation Questionnaire – Suppression subscale.
CHAPTER V
DISCUSSION

Review of Purpose

The current study thoroughly investigated the psychometric properties of scores from four measures: the SCS, the DDI, the EES, and the ERQ-S. More specifically, I conducted a series of analyses consistent with the work of Toland (2014). I calculated descriptive statistics and reliability estimates to compare values to previous research as well as provide statistics to inform future research. Next, four CFAs were necessary to determine whether the four measures met the IRT assumption of unidimensionality. Finally, IRT analyses of each of the measures involved several steps to ascertain the intricacies of each measures’ psychometric properties. It was beneficial to conduct this study despite preexisting evidence of strong psychometric properties generated via classical test theory procedures because IRT analyses allow researchers to understand how individual differences in latent trait levels influence the precision of measures.

Discussion of Results

Descriptive Statistics

The means and standard deviations of the four measures were generally comparable to those of previous research; most noticeable differences were the larger standard deviations of responses to the DDI and the EES. The greater variability of responses to the DDI and the EES suggests participants selected more extreme response
options given the Likert scale for each measure (e.g., a 1 or a 5 on the DDI). While variability in response options selected suggests a range of participant latent trait levels are represented, participants tended to select higher response options. The selection of higher response options could be beneficial for IRT analyses by allowing for higher item discriminations.

**Reliability Coefficients**

All reliability coefficients were strong and in the expected range, with the exception that the reliability coefficient for the ERQ-S was slightly greater than predicted. If the assumption of unidimensionality is met, then high reliability coefficients indicate that each item of the four measures is strongly tapping into the same latent trait (i.e., self-concealment, distress disclosure, etc.).

**Check of Assumptions**

The four CFAs generally indicated evidence of unidimensionality. The CFI and NNFI suggested the models specified for each of the four measures provided good fit to the data. However, I took caution in proceeding with the assumption of unidimensionality for the EES because of the low standardized factor loadings and marginally acceptable SRMR and RMSEA fit indices. By proceeding with caution, I acknowledge that the IRT analyses are still valid (Toland, 2014) given “adequate” unidimensionality, but that problematic items could have been removed to achieve better unidimensionality before conducting the IRT analyses.

In addition, the conclusions drawn from the standardized local dependency $\chi^2$ were mixed and inconclusive. While two measures (the SCS and the ERQ-S) showed no violations to the assumption of local independence, two measures (the DDI and the EES)
showed some evidence of local dependence. Local dependence is often problematic in IRT analyses because it can distort estimated item parameters (Toland, 2014), among other things. The “Recommendations for Future Research” below addresses how to combat such violations. So far, it appears that the EES does not strongly meet either assumption for conducting an IRT analysis.

**IRT Analyses**

The first step for each of the four IRT analyses was to check that enough participants selected each response option for every scale item (Toland, 2014). Unfortunately, this was not the case for most of the measures. With the exception of the SCS, no measure had a high enough percentage of participants select each response option for each item. It appears that each of the four measures had participants who exhibited moderate responding, or the tendency to use middle response options more frequently than extreme response options (Paulhus & Vazire, 2007). Whereas extreme responding can be problematic for normal analyses, it may have served a purpose in helping meet a minimum number of each response option for each item of each scale. In addition, lack of extreme responding suggests that many of the measures’ scales are too detailed. That is, the scales might be able to be condensed and still convey the same information as a larger Likert rating scale. I therefore collapsed across response options, which created recoded versions of the measures (e.g., I combined the DDI response options of 1 and 2 to create a recoded scale of 2 – 5). The recoded versions of the measures are thought to improve the accuracy and stability of item parameter estimates (Toland, 2014). However, I did not collect new data using the recoded scale, meaning the item calibrations could be sample specific.
Item parameters were then estimated for both item discrimination and item difficulty. Across all four measures, the lowest item discrimination was 0.87 (the EES) and the highest item discrimination was 3.85 (the DDI). Baker (2001) offers guidelines for interpreting item discrimination parameters for a logistic model. Specifically, item discrimination parameters greater than 1.70 are very high; and, as item discrimination parameters approach infinity, they perfectly discriminate (i.e., represent a step function, as described in Chapter 2). It therefore appears that approximately half of the item discrimination parameters for each scale are very high, with all of the item discrimination parameters for the DDI falling in this range. Item discrimination, or $a$, represents how strongly responding to the measure is related to the latent trait. In terms of proportion of items that met Baker’s (2001) guideline for very high item discrimination parameters, the DDI was the best, followed by the EES, the ERQ-S, and finally the SCS.

Interpretation of item difficulty parameter estimates occurs when $P(\Theta) = 0.5$. This means the item difficulty (i.e., when the probability of endorsing a dichotomy of response options compared to another dichotomy of response options is 0.5) for each item of each measure occurs at a unique theta level on the ICC. Ideally, the item difficulty parameter estimates for the original scales would be compared to the item difficulty parameter estimates for the recoded scales to determine that the recoded scales were an adequate approximation of the original scales (Toland, 2014). Due to time constraints, data collection of both the original scales and the recoded scales was not possible in the current study. Instead, item difficulty parameters were only estimated for the recoded scales. The item difficulty parameters obtained, however, suggest that all items of all measures followed a trend appropriate with IRT analyses (i.e., participants with lower
latent traits tend to endorse lower dichotomies and participants with higher latent traits tend to endorse higher dichotomies). Alone, the item difficulty parameters reveal little information about the IRT analysis.

Visual inspection of the IICs, which Toland (2014) considers a third IRT assumption, resulted in a more unanimous conclusion. The conclusion drawn from visual inspection of the IICs hinted to the needlessness of middle response options for the Likert scales: many items across all four measures had very shallow IICs for response options of 3 or 4. See Figure 9 for an example in which a response option of 3 (“neither agree nor disagree,” coded in the image as a 1) on the DDI did not achieve more item information than the other IICs at any level of theta for Item 1. A similar trend can be seen with other DDI items as well as many items of the SCS, the EES, and the ERQ-S. When each increasing response option is not more likely to be selected than previous response options as one increases along the latent trait axis, one states the response option is not “operating as expected” (Toland, 2014, p. 138). Despite the use of recoded versions of the measures, it appears the response options did not function as expected.

*Figure 9. Item information curve for Item 1 of the Distress Disclosure Index.*
Assessing IRT fit occurs at the item level and the model level. The S-$\chi^2$ statistic indicates the similarity of predicted and observed item response frequencies, with a significant value meaning poor item fit (Toland, 2014). While the SCS indicated the best item fit (i.e., only 1 item out of 10 total had poor fit), the ERQ-S indicated the worst item fit (i.e., 3 out of 4 total items had poor fit). When overall item fit is poor, Toland (2014) suggests removing the offending items, recalibrating the item parameters, and reassessing item fit. Again, due to time constraints, I was not able to follow Toland’s (2014) suggestions in the current study. In addition, the RMSEA (different from the CFA fit statistic) indicated generally good model fit; the EES was the only measure not to obtain adequate model fit. Overall, the SCS was the only measure to have consistently good item and model fit (i.e., the DDI, EES, and ERQ-S either did not have good item fit or did not have good model fit).

Finally, examination of the TICs indicated the latent trait range at which the measures excelled in predicting test information (i.e., where the TIC is greater than the standard error of the estimate curve). The larger the latent trait range, the better the measure is at predicting both low and high levels of emotional expressivity. The latent trait range averaged from theta of -1.5 to theta of +2 across measures, meaning each measure adequately captured information from participants who scored within approximately 2 standard deviations of the mean (or up to 95% of the normal distribution) of emotional expressivity.

**Limitations and Recommendations for Future Research**

Despite the advantages of IRT analyses, there are still several limitations to this study. First and foremost was the determination of what constitutes an adequate sample
size. Despite literature explaining how to conduct an IRT analysis, few authors offer
definitive recommendations as to how many subjects are necessary. Instead, texts advise
the researcher have enough responses to all response options of a measure (Embreton &
Reise, 2000) or “check that adequate numbers fall into each response category per item”
(Toland, 2014).

Second, I was unable to collect data from two different samples for each measure.
Toland (2014) states the collection of new data is necessary under several circumstances
in an IRT analysis. For one, new data should be compared to the original data if the
researcher removes items due to violations of the unidimensionality or local
independence assumptions. In addition, recoded Likert scales should be compared to the
original scales in terms of item discrimination and item difficulty parameters as well as
the inspection of the IICs. Future research should investigate the comparisons between
samples to determine if removing items helps eliminate IRT assumption violations and if
the recoded scales used in this study are justified.

Third, in addition to potentially having inadequate sample sizes, the homogeneity
of the sample limits the interpretability of the results. That is, the conclusions drawn in
this study generalize mainly to young, Caucasian women. The quality of the items (i.e.,
item and test information) could additionally differ as a function of the demographics
(e.g., age, gender, cultural values). Performing a differential item functioning (DIF)
analysis would indicate whether individuals from different demographic groups with the
same latent trait level respond to measures of emotional expressivity the same
(Embreton & Reise, 2000).
Fourth, there was dependence in the datasets. Because there were two samples but four instruments, the two sources of data inherently tie the measures together (i.e., the SCS with the EES and the DDI with the ERQ-S). Comparing data from two different samples in an IRT analysis could be dangerous if there are outliers in one sample but not the other. Ideally, one sample of participants would complete all four measures and then be analyzed using IRT. Future research might replicate these analyses among a single sample.

Finally, interpretations of the graphs generated in an IRT analysis can be subjective. Texts offer guidelines for interpreting the ICCs, IICs, and TICs of an IRT analysis, and researchers who draw the most meaningful conclusions in an IRT analysis combine graph interpretations with statistics generated from software packages. With practice, I believe interpreting IRT analyses can benefit the understanding of emotional expressivity.

**Implications**

Based on the TICs, I am comfortable with the use of all four measures being utilized for samples in which responding is not expected to be extreme. I would not suggest the use of these instruments when responding is believed to be extreme (e.g., clinical samples with very low or very high emotional expressivity). Overall, however, I do not believe researchers should have confidence when using all four measures “as-is” in practice based on my IRT results. I believe the SCS and the DDI can be used with minor alterations, whereas the EES and ERQ-S require major changes to function adequately in practice. I will explain my stipulations for improvement by measure.
First, it appears the SCS is the best functioning measure of emotional expressivity: the SCS had strong checks of the assumptions and the best response option responding percentages, suggesting a 1 – 5 Likert scale might be appropriate. The SCS was the only measure with consistent item and model fit, again indicating superiority. However, the SCS could be improved by eliminating Item 5 and removing the middle response option of “3” on the Likert scale.

My next suggestion for use in practice is the DDI because it had the best item discrimination parameters by far. Again, it would be beneficial to remove the middle response option, further reducing the DDI to a 3-point Likert scale. I also advise the elimination of several items (2, 4, 6, 9, and 11; three of which are reverse-scored) that did not obtain good item fit.

Conversely, the EES and the ERQ-S should only be utilized with major modifications. The EES did not convincingly meet either IRT assumption, meaning the good item discrimination parameters, impressive TIC, and poor model fit obtained could be artifacts of a local independence violation that generated inflated slopes, inflated scale information, and altered model-fit statistics (Toland, 2014). I suggest removing the response option of “3” on the 6-point Likert scale and rechecking whether the IRT assumptions are met before using the EES in practice.

Finally, the ERQ-S would benefit from further reduction in the Likert scale to potentially only 4 response options because the reduced range (6-point Likert scale) did not produce sufficient response options per category per item and response options of “3” and “4” did not generate substantial IICs. In addition, only one item would remain if I
eliminated items with poor item fit. Combined, my findings suggest the ERQ-S is very inadequate in terms of IRT analyses.

**Conclusions**

It is apparent from conducting an IRT analysis that classical test theory does not reveal much information about a measure. Although a reliability estimate (i.e., Cronbach’s alpha) is a good start to understanding how closely items in a measure of emotional expressivity relate to one another, it cannot reveal the intimate details of item information and test information. Classical test theory does not allow one to draw conclusions from a series of graphs and statistics that response options might need to be removed, that items might need to be eliminated, or that items might even need to be added for a measure to capture the construct of emotional expressivity. IRT analysis better informs the researcher, and practitioner, how to best implement a measure. Although it requires some knowledge and time, researchers should consider conducting more IRT analyses of emotional expressivity measures.
REFERENCES


APPENDIX

ITEM INFORMATION CURVES

Self-Concealment Scale
Distress Disclosure Index

Emotional Expressivity Scale
Emotion Regulation Questionnaire – Suppression subscale