

An Item Response Theory Analysis of Self-Report Measures of Adult Attachment

R. Chris Fraley
University of California, Davis

Niels G. Waller
Vanderbilt University

Kelly A. Brennan
State University of New York, Brockport

Self-report measures of adult attachment are typically scored in ways (e.g., averaging or summing items) that can lead to erroneous inferences about important theoretical issues, such as the degree of continuity in attachment security and the differential stability of insecure attachment patterns. To determine whether existing attachment scales suffer from scaling problems, the authors conducted an item response theory (IRT) analysis of 4 commonly used self-report inventories: Experiences in Close Relationships scales (K. A. Brennan, C. L. Clark, & P. R. Shaver, 1998), Adult Attachment Scales (N. L. Collins & S. J. Read, 1990), Relationship Styles Questionnaire (D. W. Griffin & K. Bartholomew, 1994) and J. Simpson's (1990) attachment scales. Data from 1,085 individuals were analyzed using F. Samejima's (1969) graded response model. The authors' findings indicate that commonly used attachment scales can be improved in a number of important ways. Accordingly, the authors show how IRT techniques can be used to develop new attachment scales with desirable psychometric properties.

Attachment theory is being used by an increasing number of researchers as a framework for investigating adult psychological dynamics. For instance, many researchers use this framework to study the continuity of close-relationship patterns over time (Baldwin & Fehr, 1995; Fraley, 1999; Klohnen & Bera, 1998; Scharfe & Bartholomew, 1994; Waters, Hamilton, Weinfield, & Sroufe, in press), the role of attachment organization in regulating support-seeking behavior during stressful circumstances (Fraley & Shaver, 1998; Simpson, Rholes, & Nelligan, 1992), and the influence of parents' caregiving behavior on their children's security (van IJzendoorn, 1995).

Given the diverse scope of questions addressed in attachment research, it is necessary to ensure that measures of adult attachment are as precise as possible. Until recently, however, adult attachment measures have suffered from a number of psychometric limitations (see Brennan, Clark, & Shaver, 1998; Fraley & Waller, 1998; Griffin & Bartholomew, 1994, for discussions). For example, early adult attachment instruments classified people into discrete categories (see, e.g., Bartholomew & Horowitz, 1991; Hazan & Shaver, 1987; Main & Goldwyn, 1994), oftentimes using

responses to single items to make classifications (e.g., Hazan & Shaver, 1987). Research using taxometric techniques (Meehl, 1995; Waller & Meehl, 1998) has shown, however, that adult attachment variation does not fit a taxonic model; thus, attempts to impose categorical models on attachment variability can lead to serious problems in conceptual analyses, statistical power, and measurement precision (Fraley & Waller, 1998).

More recently, researchers have focused on dimensional models of attachment (e.g., Brennan et al., 1998) and on creating multi-item inventories to assess individual differences on attachment dimensions. Despite these notable advances, it is an open question whether existing attachment scales possess the requisite psychometric properties needed to answer the diverse questions posed by attachment researchers. One limitation of many scales is that they are scored in ways that are not based on strong measurement models. For instance, an individual's degree of security is often determined by averaging (or summing) responses to statements thought to be manifestations of secure attachment. As numerous psychometricians have shown, however, such scaling techniques are problematic for many reasons (Embretson, 1996; Yen, 1986). Specifically, classical methods of scoring do not guarantee that measurement precision will be equally distributed across the domain of interest. Furthermore, the psychometric properties of so-called "total scores" depend on the number of scale items and the properties of the sample under study (Hambleton, Swaminathan, & Rogers, 1991). To avoid these problems, an explicit model is needed for relating latent variables to item response behavior.

In this article we argue that item response theory (IRT; Hambleton & Swaminathan, 1985; Lord, 1980) offers a useful

R. Chris Fraley, Department of Psychology, University of California, Davis; Niels G. Waller, Department of Psychology and Human Development, Vanderbilt University; Kelly A. Brennan, Department of Psychology, State University of New York, Brockport.

Correspondence concerning this article should be addressed to R. Chris Fraley, Department of Psychology, University of California, Davis, California 95616-8686, or to Niels G. Waller, Department of Psychology and Human Development, Peabody College, Vanderbilt University, Nashville, Tennessee 17203. Electronic mail may be sent to rcfraley@ucdavis.edu.

framework for relating latent variation in attachment organization to observed scores on self-report attachment scales. We begin by reviewing some well-known problems of classical test theory (Embretson, 1996; Waller, Tellegen, McDonald, & Lykken, 1996). Our aim in this section is to show that several limitations of classical test theory have a direct bearing on contemporary issues in attachment research, such as the degree of continuity in attachment security (Baldwin & Fehr, 1995; Klohnen & Bera, 1998; Fraley, 1999; Scharfe & Bartholomew, 1994; Waters et al., in press) and the differential stability of various attachment patterns (Davila, Burge, & Hammen, 1997). Next, we demonstrate how IRT can be used to circumvent these limitations. Specifically, we report the results of an IRT analysis of several multi-item, self-report attachment scales (Brennan et al., 1998; Collins & Read, 1990; Griffin & Bartholomew, 1994; Simpson, 1990) and discuss the limitations of these scales from an IRT perspective. Finally, we introduce two new attachment scales that were developed with the tools of IRT.

Although our primary objective is to address psychometric issues in the domain of adult attachment research, we also hope to convey the advantages of IRT-based models for other areas of personality assessment (see Reise & Waller, 1990; Steinberg & Thissen, 1995; Waller & Reise, 1989). Indeed, many of the issues of interest to attachment researchers (e.g., stability of individual differences) are relevant to researchers investigating personality and social phenomena more broadly. Unfortunately, much of the literature on IRT is too technical for many personality researchers (see any recent article in *Psychometrika* for an example). Furthermore, because many IRT models were designed for the study of intellectual abilities, social and personality researchers may find it difficult to translate traditional IRT concepts into a serviceable framework. To overcome these limitations, we have written this article in a way that is general enough to be of interest to readers specializing in diverse areas in personality and social psychology. We have also tried to make our discussion of IRT strike the delicate balance between clarity and technical thoroughness to be useful to a broad spectrum of researchers. We hope that this article will not only advance measurement in the field of adult attachment, but that it will also facilitate the exploration of IRT-based models in the broader fields of personality and social psychology.

The Basic Concepts of IRT

The rubric "item response theory" refers to a diverse family of models designed to represent the relation between an individual's item response and an underlying latent trait (van der Linden & Hambleton, 1997). In the examples considered below, we focus on IRT models for dichotomously scored items (e.g., "yes-no," "true-false") for ease of explication. However, as we explain and demonstrate later, Likert-type items (e.g., 5- or 7-point scales) can also be fruitfully analyzed by simple extensions of binary-item IRT models.

In IRT, the underlying trait is commonly designated by the Greek letter theta (θ). Theta is conceptualized as a quantitative trait and, in many IRT programs, is scaled to have a mean of zero and a standard deviation of one. An important objective in item response modeling is to characterize the relation between theta and

the probability of item endorsement. This relation, typically referred to as the *item characteristic curve* (ICC), can be defined as the (nonlinear) regression line that represents the probability of endorsing an item (or an item response category) as a function of the underlying trait. Before discussing the mathematics of an ICC, we present several illustrative ICCs in Panel A of Figure 1. There are several noteworthy features of these curves. First, they are all monotonically increasing functions. In other words, the probability of endorsing an item continuously increases as one moves along the trait continuum. This feature is typical of common parametric IRT models. Second, the curves are nonlinear. If the ICCs were linear, predicted response probabilities could be greater than one and less than zero for individuals with extreme scores. Finally, the three ICCs differ in shape. As will be discussed below, these differences are due to variability in the difficulty or extremity of the items and variability in how well the items discriminate between people with similar levels of the latent trait.

The three ICCs in Panel A of Figure 1 were generated from an item response model commonly referred to as the *two-parameter logistic item response model* (2PLM; Birnbaum, 1968):

$$P_j(\theta_i) = 1/[1 + \exp(-\alpha_j(\theta_i - \beta_j))]. \quad (1)$$

In this model, $P_j(\theta_i)$ denotes the probability that an individual with trait level θ_i will endorse item j in the keyed direction. This probability is a function of one person parameter (i.e., trait level) and two item parameters: the item difficulty and the item discrimination parameters.

In the 2PLM, the item difficulty parameter (β_j) represents the level of the latent trait necessary for an individual to have a .50 probability of endorsing the item in the keyed (i.e., trait consistent) direction.¹ For example, if an item has a difficulty value of 1.00, then an individual with a trait level of 1.00 has a 50% probability of endorsing the item. This point is illustrated by Item 1 in Panel A of Figure 1. Items with higher difficulty values tend to be endorsed only by individuals higher on the trait continuum (thus, they tend to be endorsed by few individuals). Conversely, items with lower difficulty values tend to be endorsed by individuals with moderate or high trait levels (thus, they tend to be endorsed by many individuals). This last point is well illustrated by Item 2 in Panel A of Figure 1. Notice that Item 2 has a difficulty value of -1.00 .

In the 2PLM, the item discrimination parameter (α_j) represents an item's ability to differentiate between people with contiguous trait levels. Theoretically, this parameter ranges from 0.00 to positive infinity, although in most applications the observed range is considerably shorter. Specifically, in the assessment of personality, attitudes, and interpersonal behaviors, item discriminations often fall between 0.50 and 2.50 (Gray-Little, Williams, & Hancock, 1997; Hambleton et al., 1991; Kim & Pilkonis, 1999; Reise

¹ The term *difficulty* is used because many psychometric models, including IRT-based models, have been developed for the assessment of educational abilities (Crocker & Algina, 1991). In the context of IRT, an item is considered difficult if a high level of ability or knowledge is required to answer it correctly. Only individuals with a high degree of knowledge will be able to answer the difficult items, and almost everyone will be able to answer the easy items.

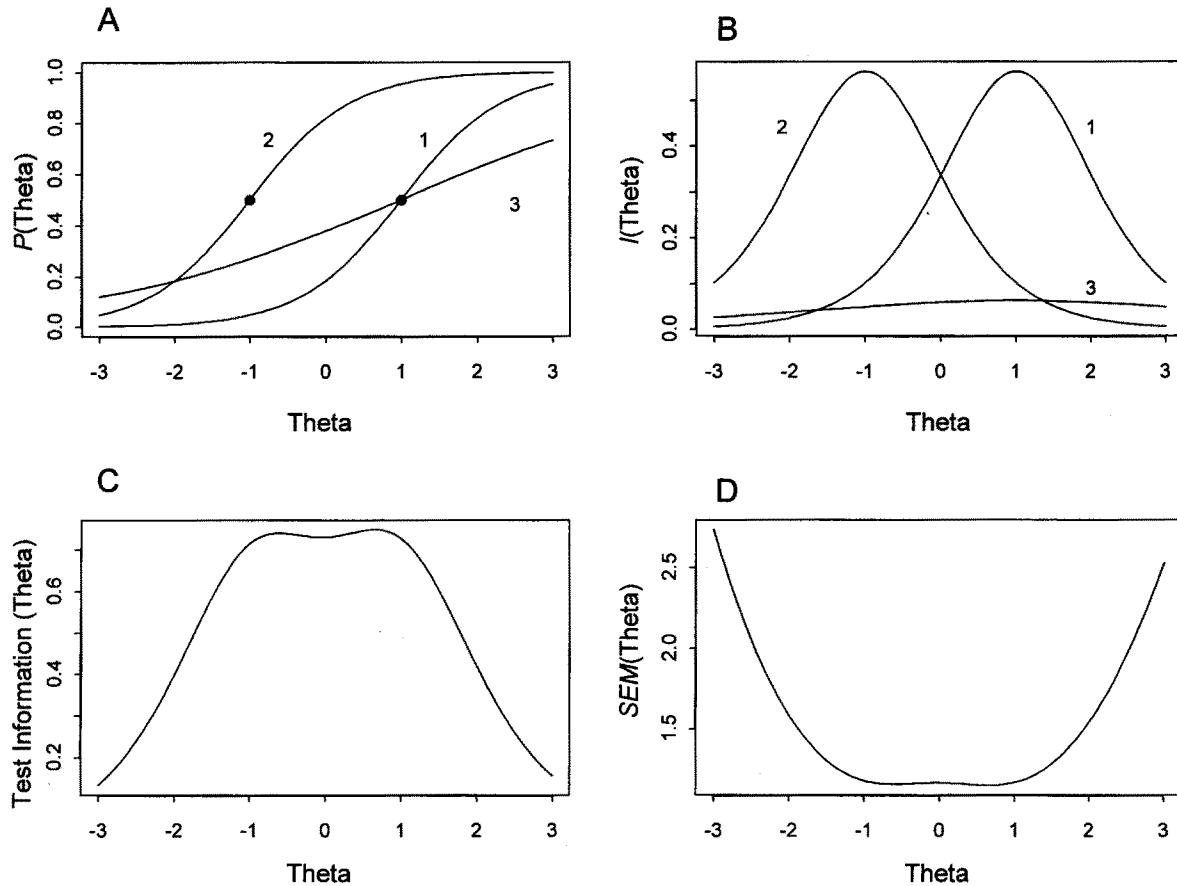


Figure 1. Item response properties for a three-item hypothetical scale, using the two-parameter logistic item response model (2PLM). A: item characteristic curves for the three items. B: information curves for each item. C: test information function for the three-item scale. D: standard error of the test as a function of the latent trait.

& Waller, 1990). When the latent trait distribution is normal, the item discrimination parameter is related to the item factor loading (when the factor analysis has been conducted on tetrachoric correlations; Takane & de Leeuw, 1987). It is also related to the well-known item–test correlation of classical test theory. Accordingly, an item that correlates highly with a scale total score is a better indicator of the latent trait than an item that correlates less strongly with the total score. Similarly, in IRT, an item that has a high discrimination value is a better indicator of the latent trait than an item that has a smaller discrimination value.

To better understand the above points, consider, once again, Item 1 in Panel A, Figure 1. This item has a discrimination value (α) of 1.50 (which roughly corresponds to a factor loading of .83). Notice that people with trait levels in the vicinity of 1.25 are substantially more likely to endorse Item 1 than are people with trait levels of 0.75. Item 3, however, has a discrimination value of only 0.50 (roughly corresponding to a factor loading of .45). Notice that in this case, individuals with trait levels near 1.25 are only slightly more likely to endorse Item 3 than are persons with trait levels of 0.75. In other words, Item 3 does a poor job of discriminating among individuals with similar trait levels.

It is important to note that an item's ability to discriminate between people with similar trait levels is highest in the theta region corresponding to the item's difficulty. For example, Item 1—which has a difficulty value of 1.00—is better able to differentiate between people with scores of 0.75 and 1.25 than between people with scores of -1.25 and -0.75 . This observation has an important implication for test construction and evaluation within an IRT framework. Namely, it suggests that items are not equally informative across the entire trait range. Some items are adept at discriminating people on the high end of the trait continuum but poor at discriminating people in the lower range of the continuum.

In IRT, these relations are represented by the concept of item information (I). Formally, item information in the 2PLM is defined as:

$$I_j(\theta_i) = \alpha_j^2 \times P_j(\theta_i) \times (1 - P_j(\theta_i)), \quad (2)$$

where α_j^2 is the squared item discrimination parameter for item j and $P_j(\theta_i)$ is the probability of endorsing item j for individuals with θ level i . According to this equation, an item yields the most information at the point on the trait continuum where θ_i equals β_j .

In other words, items are most informative when the item difficulty parameter is perfectly matched to the person's trait level.

When item information is calculated for all trait values, we can construct an item information curve. Like ICCs, item information curves (IICs) can be plotted to represent relative information as a function of trait level. For instance, in Panel B of Figure 1 we have plotted item information functions for the three items discussed previously. Notice that these plots graphically portray our earlier point. Namely, an item is most informative at trait or theta levels corresponding to the item difficulty values. Furthermore, as implied by Equation 2, items with small discrimination values (e.g., Item 3) yield relatively less information than items with high discrimination values. Such items have ICCs that are flatter, or less peaked, than items with high discrimination values.

Importantly, item information curves can be summed to produce an information curve for the full scale. This curve is often called the test information curve (TIC). The TIC represents the relative precision of the scale across different levels of the trait continuum, and the height of the TIC is proportional to the standard error of measurement (*SEM*). Specifically, in the 2PLM, the standard error of a trait estimate is equal to the inverse square root of the information value at a particular trait level. In IRT models, measurement precision can potentially differ for people with different trait levels. Unlike classical test theory, in which measurement precision is typically represented by a single number (such as Cronbach's alpha), in IRT there are as many standard errors of measurement as there are unique trait estimates. As one might suspect, some tests, like items, are particularly well-suited to specific regions of the trait domain.

The TIC and the conditional standard error of measurement for the three illustrative items of this section are presented in Panels C and D of Figure 1. As can be seen in these plots, this three-item scale provides the most precise measurement for individuals with trait values between -1.00 and $+1.00$. Notice that people on the extreme high and low ends of the trait distribution are measured with relatively less precision.

Advantages of IRT Over Classical Scaling Methods

A major advantage of IRT models is that they are based on an explicit measurement model that characterizes the relation between a latent trait and an observable manifestation of the trait. In other words, IRT is a model-based approach to psychological assessment (Embretson, 1996; Embretson & Hershberger, 1999; Waller et al., 1996; Zickar, 1998). Because of this feature, IRT offers a number of advantages that traditional methods based on classical test theory do not. We will explain one such advantage below and discuss how it is relevant to current theoretical issues and debates in adult attachment research. (See Hambleton et al., 1991, for further discussion of the advantages of IRT over classical test theory.)

As we noted previously, a major limitation of traditional assessment frameworks is the assumption that measurement precision is constant across the entire trait range. IRT models, however, explicitly recognize that measurement precision may not be constant for all people. It is probably the case that most scales, particularly those derived for clinical purposes, are good at differentiating among people in the high range of a trait, but less well-suited for

differentiating people in the normal-to-low range of the trait. In other words, it is likely that many scales used in personality research have an unequal distribution of precision across the normal range of the trait continuum.

A scale can have uneven information functions for at least two reasons: (a) The item difficulty values cluster in a narrow region of the trait range, and/or (b) the item discrimination values are differentially concentrated in certain regions of the trait range. Below, we focus on uneven scale information resulting from the first of these sources and highlight its implications for investigating two issues in contemporary attachment research: (a) the stability of individual differences in security and (b) the differential stability of anxious attachment.

When a scale has an uneven distribution of item difficulty values, it also will have an uneven information function. For instance, when the item difficulty values are clustered on the high end of the trait range, the scale will yield precise measurement for individuals with high traits values and relatively imprecise measurement for individuals with moderate to low trait values. To illustrate the psychometric properties of such scales, we created two hypothetical 20-item scales that differed in their distribution of item difficulty values. For both scales, all items had discrimination values of 1.50. In the first scale, which we will call the "clustered scale," the 20 items had evenly spaced difficulty values that ranged from 1.00 to 3.00. In the second scale, which we will call the "nonclustered scale," the difficulty values spanned a wider range. Specifically, the items had evenly spaced difficulty values between -3.00 and 3.00 (an interval that essentially covers the entire trait range). In a simulation, we created normally distributed, error-free latent trait values for a sample of 200 people ($M = 0.00$, $SD = 1.00$) and, by applying Equation 1, used these values to generate expected observed scores on the two scales.

The test information curves for the clustered and nonclustered scales are illustrated in panel A of Figure 2. Notice that the nonclustered scale (labeled *N* in the figure) has an information curve that is relatively constant across a wide range (-2.00 to 2.00) of the trait continuum. The clustered scale (labeled *C* in the figure), however, has an information curve that peaks on the high end of the trait range and tapers off toward the lower end of the trait continuum. The hill-shaped feature of this information curve nicely illustrates why scales with clustered item difficulties provide uneven measurement precision across the trait range. An important feature of these scales is that they both have Cronbach alpha reliabilities of .81 in the simulations. Nonetheless, as clearly indicated in the figure, the relative degree of measurement precision for individuals at the low and high ends of the trait continuum are dramatically different for the two scales. For instance, the conditional standard error of measurement for individuals with trait values of -1.00 are 4.77 and 0.49 for the clustered and nonclustered scales, respectively.

What are the implications of clustered item difficulties for estimates of personality stability? To examine this issue, we conducted a second simulation in which several parameters were varied. Specifically, we manipulated the degree of "true" trait stability by varying the cross-time correlation, r , from $r = 0.00$ (no stability) to $r = 1.00$ (perfect stability) in increments of 0.10. For each level of stability, we generated two sets of 200 z scores from a bivariate normal distribution to represent the error-free latent

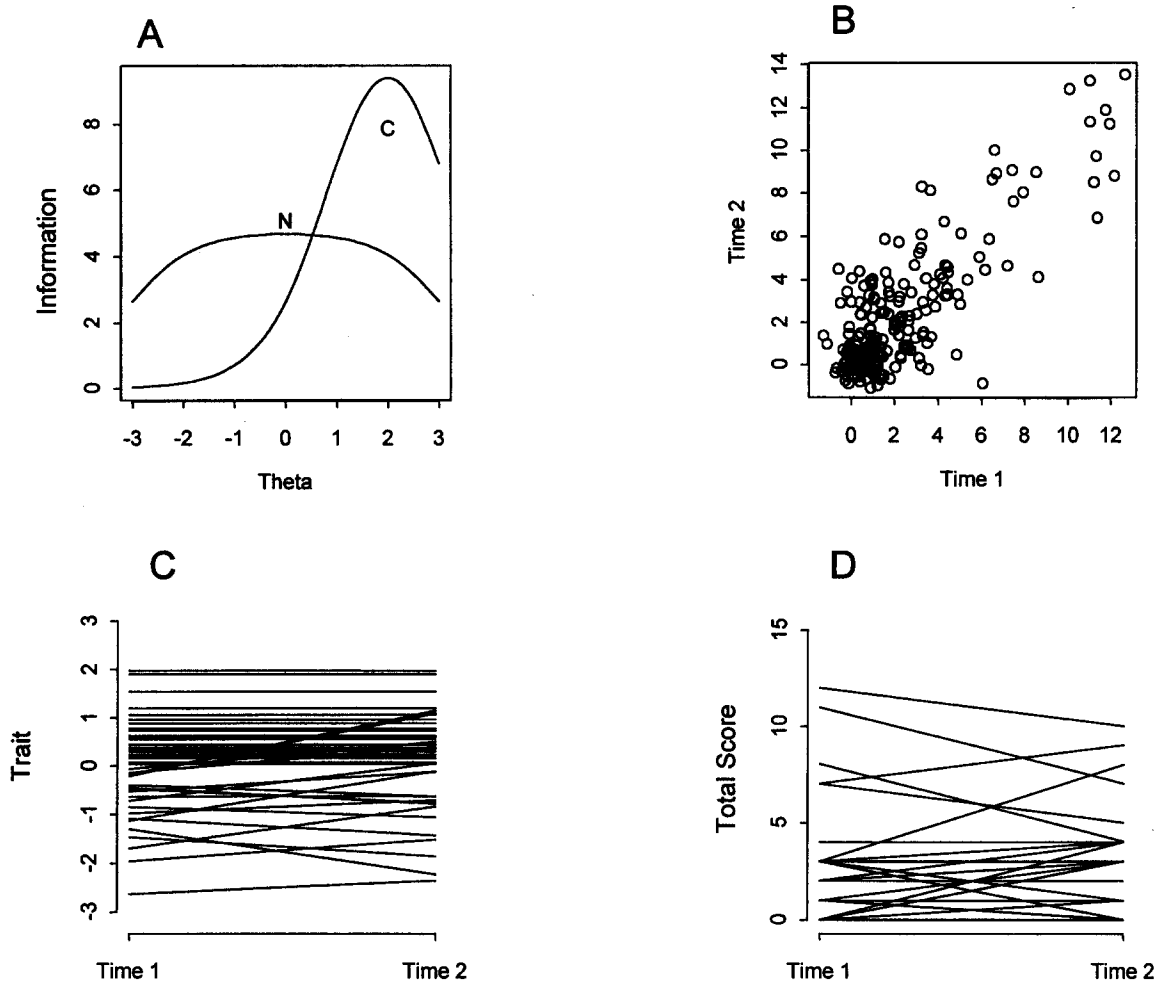


Figure 2. Item response properties for scales with clustered and evenly distributed item difficulty values. A: information curves for both kinds of scales. B: scatterplot for Time 1 and Time 2 observed scores for a scale with clustered difficulty values. C: growth curves for latent trait values in which people high in the trait are perfectly stable over time and people low in the trait are less stable. D: growth curves for the clustered-scale observed scores based on the latent trait scores in Panel C.

trait values. We then generated Time 1 and Time 2 item responses for the clustered and nonclustered scales that were described previously. The mean test-retest correlations for both scale types, averaged across 500 simulations, are presented in Table 1.

Notice that the clustered scale provides trait stability estimates that are similar to those of the nonclustered scale across all levels of test-retest stability. What is not evident from these results is that the test-retest correlations for the clustered scale are largely an artifact of the distribution of item difficulties. To clarify this point, we have plotted Time 1 versus Time 2 total scores for the clustered scale in Panel B of Figure 2. Notice that there is a high density of observations at the bottom left-hand side of the plot. The high density of scores in this region occurs because the test items are too difficult for most people in the sample. In other words, a high level of the latent trait is required for an individual to endorse the items in the keyed direction. Few people have the necessary trait level to yield a keyed response; consequently, most people fail to

endorse the items and consistently receive low observed scores at each time point. This result, combined with the fact that people with high trait values are assessed with considerable precision (see Panel A), is responsible for the surprisingly high test-retest correlation in a scale with markedly uneven measurement precision. An important implication of these findings is that high test-retest correlations can be obtained in situations where measurement precision is uneven at both time points.²

Scales with clustered difficulty values present an additional problem that, potentially, can vitiate developmental research on

² As mentioned previously, uneven measurement precision can also result from having a nonuniform distribution of discrimination values. When this is the case, test-retest estimates of reliability are substantially attenuated relative to cases in which the item discrimination values are constant across the entire trait range.

Table 1
*The Effect of Clustered Item Difficulties on Test-Retest
 Estimates of Trait Continuity*

True degree of continuity	Test-retest correlation	
	Nonclustered scale	Clustered scale
.00	.00	.00
.10	.08	.07
.20	.16	.13
.30	.25	.21
.40	.33	.28
.50	.41	.36
.60	.49	.44
.70	.57	.53
.80	.66	.62
.90	.74	.72
1.00	.82	.82

Note. These statistics are based on the average results of 500 simulations.

attachment. For instance, in Panels C and D of Figure 2, we have plotted simulated scores on two administrations of the aforementioned clustered scale. For illustrative purposes, we will assume that our scales measure individual differences in attachment-related anxiety. In Panel C, we display the latent trait (i.e., true trait) values for the two administrations. The data were generated to conform to the following model: Individuals with trait values greater than 0.00 at Time 1 were assumed to have perfectly stable trait values; individuals with trait values less than or equal to 0.00 at Time 1 were assumed to undergo change. In particular, for these individuals, the correlation between their Time 1 and Time 2 trait values was precisely .50. The model design is easily discerned in the linear growth curves plotted in Panel C of Figure 2. Notice that Time 1 latent trait values greater than 0.00 do not change and, hence, the corresponding linear growth curves are horizontal (i.e., have slopes of 0.00). However, individuals with Time 1 latent trait values that are less than 0.00 show considerable change and have linear growth curves with nonzero slopes.

Investigators with access to the (true) latent trait values that are portrayed in Panel C would accurately conclude that the stability of attachment-related anxiety is a function of one's trait value. Highly anxious individuals show little-to-no change, whereas relatively less anxious individuals show considerable change. Unfortunately, except in simulation studies, we never have access to actual latent trait values—at best, we obtain estimated latent trait values. Most researchers use estimated trait scores that are simple sums (or means) of observed item responses (i.e., total scores). When scales have psychometric properties that are similar to those of the so-called "clustered scale" of our running example, this practice can provide grossly misleading results. This point is illustrated in Panel D of Figure 2. In this panel we have plotted the Time 1 and Time 2 observed or total scores for the same individuals that were used to construct Panel C. Recall that individuals with high latent trait values did not change whereas individuals with low latent trait scores exhibited considerable change. At the observed score level, however, the pattern of results is exactly opposite of that of the latent trait level. The observed score growth curves suggest that highly anxious individuals are less stable over

time than are less anxious individuals. In fact, this later observation is precisely what some researchers have reported. For instance, Davila and her colleagues recently found that highly anxious individuals (i.e., those people who are anxiously concerned about their partner's responsiveness and availability) exhibit less stability in their attachment patterns than less anxious people (Davila et al., 1997). Without knowing the information properties of self-reported anxiety scales, however, it is difficult, if not impossible, to know what these observations reveal about differential stability at the latent trait level.

An IRT Analysis of Adult Attachment Items

The many limitations of classical test methods for studying issues relevant to adult attachment theory make it imperative to determine whether existing attachment scales are adequate from an IRT perspective. Specifically, it is necessary to ascertain whether existing scales have a high and evenly distributed degree of measurement precision. To this end, we conducted an IRT analysis of four commonly used multi-item self-report inventories of attachment: Brennan et al.'s (1998) Experiences in Close Relationships scales, Collins and Read's (1990) Adult Attachment Scales, Griffin and Bartholomew's (1994) Relationship Styles Questionnaire, and Simpson's (1990) (unnamed) attachment scales. In the sections that follow, we show that all of these scales are less than ideal from an IRT perspective. In particular, most of the scales have relatively low or unevenly distributed test information curves. Thus, measurement precision is either poor or differentially distributed across the trait range with these scales. Some of these limitations are easily avoided, however, and we demonstrate how IRT methods can be used to develop new scales with improved psychometric properties.

Method

Sample and Instruments

The data used in this study were originally collected by Brennan and her colleagues (Brennan et al., 1998). The sample contains item responses from 1,085 undergraduate students (682 women, 403 men) from the University of Texas at Austin. At the time of testing, the participants had a median age of 18 years (range = 16–50). Further information concerning sample characteristics is reported elsewhere (Brennan et al., 1998).

All participants were administered 323 items designed to measure attachment organization. All items were rated on a 1 (*strongly disagree*) to 7 (*strongly agree*) Likert-type scale and were worded to be relevant to romantic relationships. The items were drawn from 14 self-report inventories of attachment that were available in 1996 (see Brennan et al., 1998, for a listing of inventories). Included in this diverse item pool were items from the four inventories that we focus on here: Brennan et al.'s (1998) Experiences in Close Relationships Questionnaire (ECR); Collins and Read's (1990) Adult Attachment Scale (AAS); Griffin and Bartholomew's (1994) Relationship Styles Questionnaire (RSQ); and Simpson's (1990) (unnamed) attachment questionnaire.

The ECR assesses two dimensions: Anxiety and Avoidance. An 18-item subscale measures each dimension. The AAS assesses three dimensions: Close, Depend, and Anxiety. Six items are used to assess each dimension. The RSQ measures a person's relative fit to four theoretical attachment types: Secure, Fearful, Preoccupied, and Dismissing. The inventory consists of four subscales, with 4 to 5 items each. The Simpson inventory assesses people's relative fit to three attachment types: Secure, Avoidant,

and Anxious attachment. Each subscale consists of 4 to 5 items. For a detailed discussion of these various subscales and their theoretical interpretations, see Brennan et al. (1998); Crowell, Fraley, and Shaver (1999); and Fraley and Shaver (in press).

The Graded Response Model (GRM)

The item parameters of Samejima's (1969, 1996) GRM were estimated for each 7-point Likert item used in this study. The GRM is a potentially useful item response model when item response options can be conceptualized as ordered categories (e.g., with Likert-type rating scales). Samejima's model is an extension of the 2PLM for dichotomous item responses discussed previously.

Within the GRM framework, an item response scale is conceptualized as a series of $m - 1$ response dichotomies, where m represents the number of response options for a given item. Thus, an item rated on a 1-to-4 scale has three response dichotomies: (a) Category 1 versus Categories 2, 3, and 4; (b) Categories 1 and 2 versus Categories 3 and 4; and (c) Categories 1, 2, and 3 versus Category 4. In now classic work, Samejima (1969) showed that when an item, i , is conceptualized as a series of ordered dichotomous response options, the 2PLM can be generalized to estimate the response option probabilities. For instance, Samejima's model considers the probability of endorsing each response option category, x_j , or higher as a function of a latent trait [$P_{x_j}(\theta_i)$]. For the hypothetical 4-point item, we can generate three ICCs.³ The response option difficulty represents the point on the latent trait continuum where there is a 50% chance of endorsing the x_j or higher response option. In this sense, the response option difficulty represents a between-option "threshold" parameter. For each item, the number of threshold values equals the number of response dichotomies ($m - 1$). In the GRM, each item has a single discrimination (α) value for all response options.

The ICCs for each response dichotomy can be used to calculate the probability of endorsing a particular response option, x_j , as a function of the latent trait. These probability functions are known as category response curves (CRCs; Embretson & Reise, in press). Once the ICCs for each response dichotomy are known, the CRC for a particular response option, x_j , is given by the following equation:

$$P_{x_j}^*(\theta_i) = P_{x_j}(\theta_i) - P_{x_{j+1}}(\theta_i), \quad (3)$$

where $P_{x_j}(\theta_i)$ is the probability of endorsing option x_j or higher and $P_{x_{j+1}}(\theta_i)$ is the probability of endorsing the next highest option, $x_j + 1$, or higher. The probability of endorsing the lowest response category or higher [$P_{x_1}(\theta_i)$, in this example] is 1.00, by definition. Similarly, the probability of endorsing category $m + 1$ [$P_{x_m}(\theta_i)$, in this example] or higher is necessarily 0.00 because this response option does not exist. Thus, with a 4-point scale, the four possible CRCs are given by the following:

$$P_{1j}^*(\theta_i) = P_{1j}(\theta_i) - P_{2j}(\theta_i),$$

$$P_{2j}^*(\theta_i) = P_{2j}(\theta_i) - P_{3j}(\theta_i),$$

and

$$P_{3j}^*(\theta_i) = P_{3j}(\theta_i) - P_{4j}(\theta_i),$$

$$P_{4j}^*(\theta_i) = P_{4j}(\theta_i) - P_{5j}(\theta_i),$$

where $P_{1j}(\theta_i) = 1.00$ and $P_{5j}(\theta_i) = 0.00$.

To better illustrate these points, several example ICCs and CRCs for a 4-point item that fits the GRM are given in Panels A and B of Figure 3. The example item has a discrimination value of 1.5 and difficulty or threshold values of -1.50 , 0.00 , and 1.50 . Notice that in the GRM, in contrast to the 2PLM discussed previously, the CRCs are not necessarily monotonically increasing functions of the latent trait values. For example, as illustrated in Panel B of the figure, the probability of endorsing Response Option 2 (or 3) increases as one moves from the low to the mid-range of the trait

distribution, at which point the probability begins to decrease. As with dichotomous items, item information functions can be generated for graded response items. Equations for the GRM information functions are provided in Samejima (1969, Equation 6-6, p. 39).

In the GRM, as in the 2PLM, scale information is obtained by summing the item information functions. Similarly, the conditional standard error of measurement for a given trait value equals the inverse square root of the information level at that theta value. Item information and standard error functions for our example 4-point item are illustrated in Panels C and D, respectively, of Figure 3. As can be seen in these panels, this item does a relatively good job at distributing its precision equally across trait regions between -1.5 and $+1.5$. The overall degree of precision is rather low, however, because we are considering only a single item.

Results

Estimating GRM Item Parameters for Existing Attachment Scales

Our first objective was to determine the psychometric properties of the 12 scales from four well-known attachment inventories—the ECR, AAS, RSQ, and the Simpson inventory (Brennan et al., 1998; Collins & Read, 1990; Griffin & Bartholomew, 1994; Simpson, 1990) from an IRT perspective. To achieve this goal, we estimated GRM item parameters for all items within each subscale of the four inventories (85 items total). All IRT analyses were conducted with Multilog Version 6.0 (Thissen, 1991), a program designed to estimate a wide variety of item response models. We estimated seven parameters for each item: one item-discrimination value and six item-difficulty or between-category threshold values.

A testable assumption of the GRM is that item covariation arises predominantly from a single underlying dimension (i.e., the unidimensionality assumption). A common way to test this assumption is to examine the ordered eigenvalues from the item correlation matrix (see Hambleton et al., 1991, chap. 5). When the unidimensionality assumption is tenable for a subscale, the first eigenvalue should be considerably larger than the remaining eigenvalues. A dominant first dimension was found for all scales except Griffin and Bartholomew's (1994) Secure and Preoccupied scales. These scales have only four to five items each; thus, these findings are not unexpected. (A table of ordered eigenvalues is available upon request from the authors.)

There are many methods for assessing model-data fit in an IRT analysis (Drasgow, Levine, Tsien, Williams, & Mead, 1995). For our analyses, we focused on two useful methods: (a) item residual plots (Hambleton et al., 1991, chap. 5) and standardized weighted mean squares of the item residuals (Wright & Masters, 1982, chap. 5). Item residual plots provide a visual depiction of the discrepancies between the predicted response frequencies for each response category, given the estimated model, and the observed response frequencies for each cate-

³ We have retained the acronym ICC in order to remind the reader that these probability functions are calculated from the 2PLM discussed previously. However, it is important to keep in mind that these curves are not item characteristic curves per se in the context of the GRM because several such curves exist for each response dichotomy within an item.

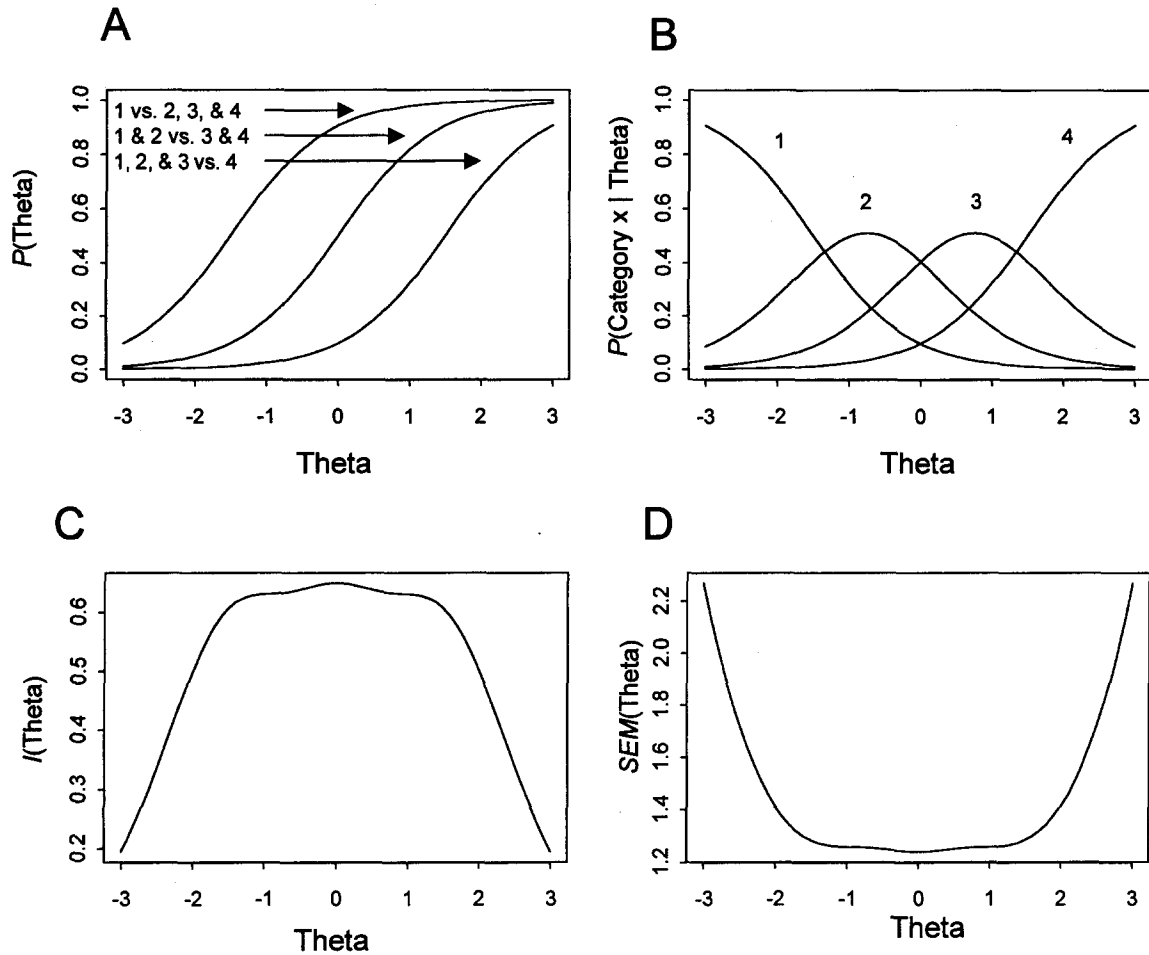


Figure 3. Item response properties for an item rated on a 1- to 4-point Likert-type scale. A: characteristic curves for each response dichotomy. B: category response curves for each response option. C: information for the item. D: standard error for the item as a function of θ .

gory. We constructed item residual plots for the 85 items in this analysis and inspected them for signs of model misfit. Virtually all of the items showed evidence of good fit. We also examined the standardized weighted mean squares of the item residuals based on a method described by Wright and Masters (1982, p. 100). In short, this method allows one to compute a quantitative index of the discrepancy between the observed responses and the model implied responses. These discrepancies are standardized and weighted as a function of the variances of the expected response probabilities. An examination of these indexes indicated that all of the items showed evidence of good fit.

Figure 4 displays the test information functions for the 12 attachment scales. As can be seen in these plots, most of the curves are relatively low. This indicates that the overall degree of measurement precision for these scales is also relatively low. Despite this limitation, most of the scales have relatively uniform measurement precision across wide regions of their respective trait ranges. A noteworthy exception to this general pattern is found in the "secure" region of the scales. Specifically, the scales are less

precise for measuring individuals with θ levels falling above 1.00 on scales keyed toward security (e.g., the AAS Close scale) and below -1.00 on scales keyed toward insecurity (e.g., Simpson's Avoidant scale). Perhaps the most striking feature of these plots is the relatively greater degree of measurement precision provided by the ECR scales (shown in the upper leftmost panels of the figure).

Using IRT to Construct Attachment Scales With Improved Psychometric Properties

In the aforementioned analyses, we found that the ECR scales had test information functions that were clearly higher than those of the other attachment scales. This observation suggests that they may be preferable to the alternatives. However, it may be possible to construct better scales for measuring Anxiety and Avoidance by using IRT techniques. Recall that the ECR scales, like the other scales, were unable to assess the secure end of each dimension with the same degree of fidelity as the insecure end. By explicitly

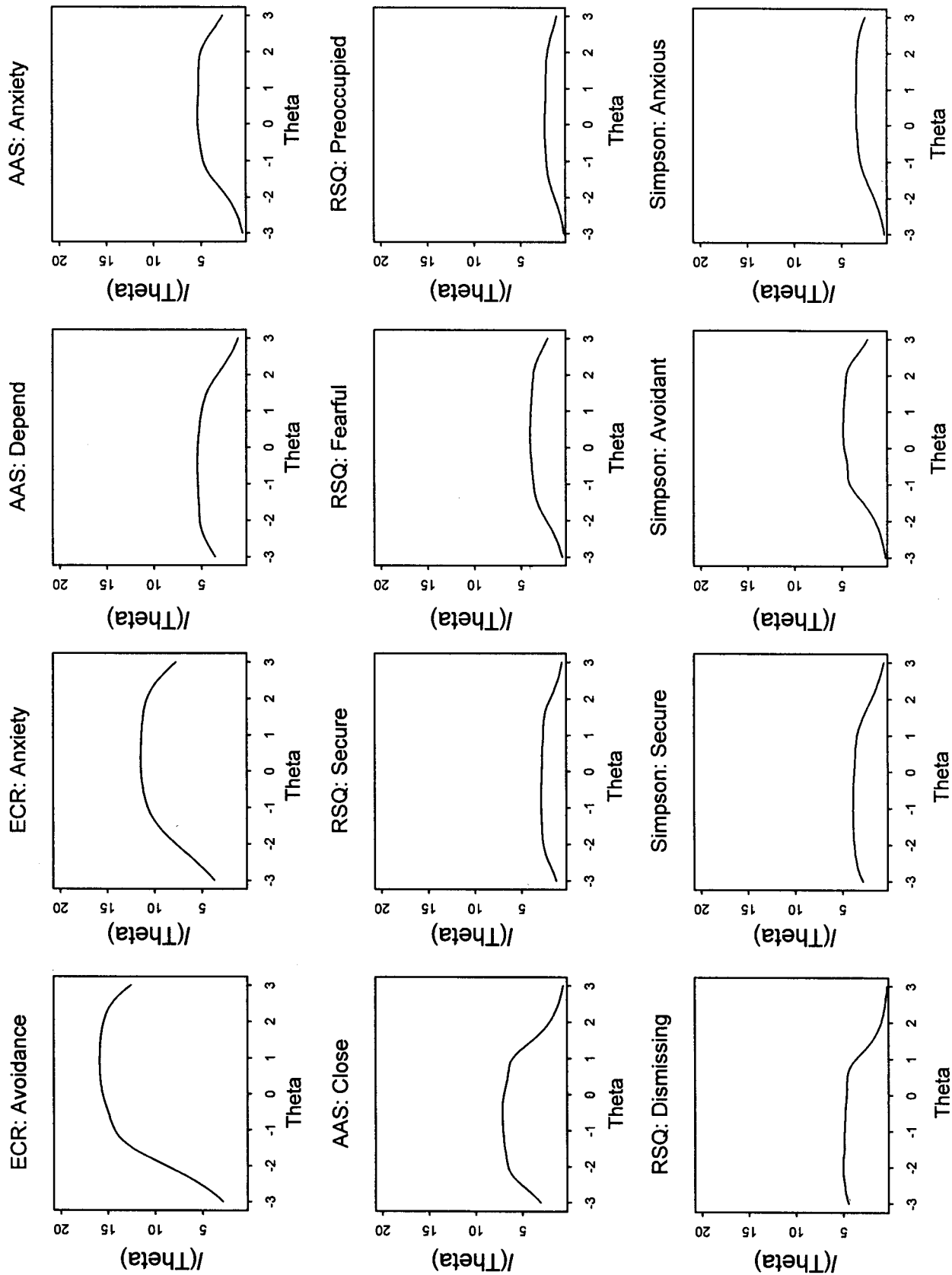


Figure 4. Test information curves for each attachment subscale.

studying the item difficulty or threshold values, especially those falling on the low end of the trait range (i.e., the β_1 values), we may be able to construct scales that assess the low end of each attachment dimension with the same degree of precision as the middle to high end. Furthermore, by attending to item discrimination values, it should be possible to create scales that have more precision than the original ECR scales—without increasing the number of scale items.

To explore these possibilities, we turned our attention to the complete pool of 323 items collected by Brennan et al. (1998). This diverse data set is well suited for exploring variation in item properties because it contains a wide array of items drawn from various theoretical approaches to attachment. To identify the purest markers for Anxiety and Avoidance, the theoretical dimensions that the ECR is designed to capture (see Brennan et al., 1998; Crowell et al., 1999), we performed a principal-axis factor analysis on 30 clusters of homogeneous items derived from a cluster analysis of the full 323 item pool. There are two advantages to factoring clusters of items rather than individual items. First, individual items are less reliable than item clusters. Second, in a factor analysis, factors are partly defined by the number of items representing each factor domain. Because there is some degree of variability in the theoretical orientation of the various scales and, hence, in the number of items contained in each inventory, a factor analysis of individual items might give differential weight to the theoretical constructs embodied by longer inventories. By aggregating responses to homogeneous items in a single index, the resulting factors are more likely to be defined by theoretical content rather than item frequency.

We clustered the 323 items using the Partitioning Around Medoids (PAM) clustering algorithm (see Kaufman & Rousseeuw, 1990, chapter 2) that is included in the S-Plus programming language (MathSoft, 1999). Like the more familiar *k*-means clustering procedure (Hartigan & Wong, 1979), PAM allows the user to select the desired number of clusters. However, unlike *k*-means clustering, PAM is a robust technique that does not require user-selected starting values. After examining a wide range of cluster numbers (i.e., 20–32 clusters), we concluded that a 30-cluster solution provided the best partitioning of items for our purposes. Specifically, each cluster contained a set of conceptually tight items that we judged to be sufficiently different in content from the remaining clusters. The content of the 30 clusters can be succinctly summarized by the following cluster labels: (1) *anxiety about abandonment*, (2) *fear of intimacy*, (3) *no anxiety about abandonment*, (4) *desire to merge*, (5) *I drive others away*, (6) *dependency/preoccupation*, (7) *don't depend on others or express emotions*, (8) *fear of rejection/relationships are risky*, (9) *I prefer distance*, (10) *open communication*, (11) *I'm important*, (12) *I can't trust others*, (13) *dismissing*, (14) *I value independence*, (15) *I fear disapproval*, (16) *anger/frustration*, (17) *partner not sensitive*, (18) *can't depend*, (19) *desire to be closer*, (20) *partner unpredictable*, (21) *values achievement*, (22) *easy to be close*, (23) *partner available*, (24) *ambivalence*, (25) *I want to be nearby my partner*, (26) *I'm not lovable*, (27) *I'm lovable*, (28) *preoccupied*, (29) *people are good*, and (30) *partner is sensitive*. We created cluster scores for each of the 30 clusters by averaging peoples' responses to the items within each cluster.

We next conducted a principal-axis factor analysis on the 30 cluster scores, followed by varimax rotation. Because we were seeking to construct improved scales for assessing the dimensions of Anxiety and Avoidance, we examined the first 2 factors of a 3-factor solution.⁴ Interestingly, inspection of the factor pattern matrix revealed that the clusters tended to fall along the perimeter of a hypothetical circle. To illustrate this characteristic of the solution, we have plotted the location of the 30 clusters in the 2-dimensional factor space in Figure 5. The circular pattern of factor loadings indicates that there is no simple structure in the data, and, consequently, that the particular rotation obtained by varimax is arbitrary. Although the transformation did find a rotation that maximized the varimax criterion (i.e., the sum of the variance of the squared factor loadings across factors), the differences between the sums obtained by the varimax rotation and other possible rotations were negligible. Therefore, guided by theoretical considerations (Brennan et al., 1998), we manually rotated the axes counterclockwise 70 degrees such that the first factor was aligned with clusters focused on attachment-related anxiety (e.g., separation and rejection anxiety), and the second factor was aligned with clusters representing avoidance (see Figure 5). After rotating the axes to these theoretical targets, we used the rotated factor loadings to generate weighted least squares factor score estimates for each individual.

To obtain independent markers of each dimension for the IRT analyses, we selected items that correlated higher than .40 with scores on one factor (e.g., Avoidance) and less than .25 with scores on the other factor (e.g., Anxiety). Sixty-seven items met this criterion for Anxiety; 78 items met this criterion for Avoidance. Because many of the items overlapped considerably with respect to item content, we removed items that we judged to be blatantly redundant. As a result, 40 items remained for Anxiety and 50 for Avoidance.

GRM item-parameters items were estimated separately for the 40 Anxiety items and the 50 Avoidance items. The unidimensionality assumption for each item set appeared warranted: The ordering of the first three eigenvalues for the Anxiety items was 12.13, 2.10, and 1.99; the ordering of the first three eigenvalues for the Avoidance items was 17.20, 2.85, and 1.99. An examination of item residual plots and standardized weighted mean squares of the item residuals indicated that the item responses were well modeled by the GRM, with the exception of 1 Anxiety item. Thus, we removed this item from the pool and reestimated the parameters for the remaining 39 Anxiety items.

Recall that our primary goal was to construct scales for assessing Anxiety and Avoidance that would possess both a high and uniform degree of information. Unfortunately, the IRT analyses revealed that our item pool did not contain many items capable of assessing the low end of the Anxiety and Avoidance dimensions well. The median β_1 value, for example, was -1.67 for Anxiety and -1.86 for the Avoidance items. Furthermore, the items that did have low β_1 values tended to have low discrimination values

⁴ As discussed by Wood, Tataryn, and Gorsuch (1996), it is desirable to extract one factor greater than the hypothesized number of factors to reduce error in the estimated factor loadings of interest.

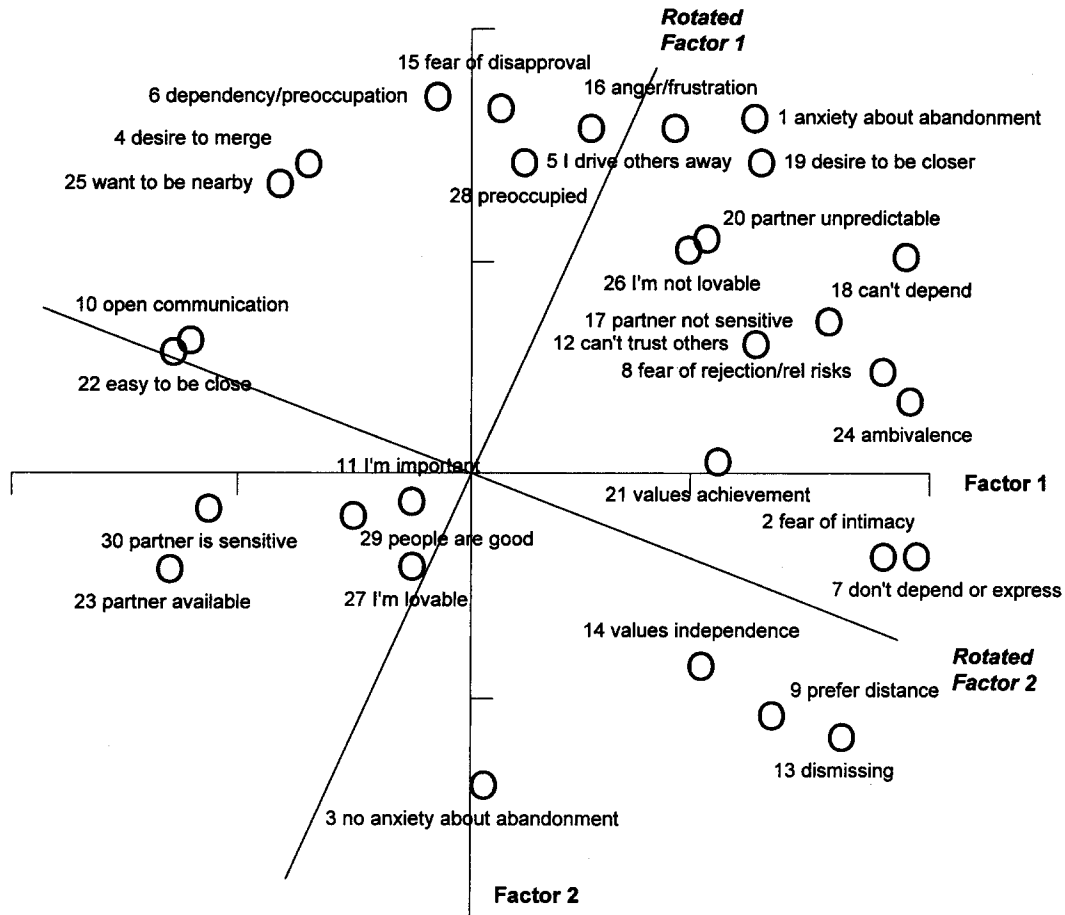


Figure 5. Factor loading plot for the 30 attachment clusters.

as well (the correlation between α and β_1 was .59 for the Anxiety items and .68 for the Avoidance items). In other words, the properties of the items in the item pool prevented us from creating scales that simultaneously covered a wide trait range and had a high degree of precision. Given this constraint, we decided to select items on the basis of their discrimination values alone. For each scale, we chose the 18 items with the highest discrimination values. Thirteen of the 18 Anxiety items (72%) were in the original ECR Anxiety scale. Seven of the 18 Avoidance items (39%) were in the original ECR Avoidance scale. Because there is some degree of overlap between the new items and the original ECR items, we refer to these two new 18-item scales as the Experiences in Close Relationships Questionnaire—Revised (ECR-R). The ECR-R items and their estimated parameters are presented in Tables 2 and 3.

The test information functions for the two ECR-R scales are illustrated in Figure 6.⁵ For comparison, we have superimposed the test information functions for the original ECR scales. Notice that the items selected on the basis of IRT techniques contain a substantially higher degree of information than the original scales. In fact, the TIC for the revised Anxiety scale is almost twice as high as the TIC for the original Anxiety scale. However, also notice that

the two ECR-R scales, like the original ECR scales, are not adept at assessing individuals with trait levels less than -1.00 on Anxiety or Avoidance.

Although these scales improve on the original ECR scales, we believe that there are some limitations to the ECR-R that need to be made explicit. First, the ECR-R scales, like the other attachment scales examined here, assess high levels of security (i.e., low Anxiety and low Avoidance) with considerably less precision than insecurity. Ultimately, this stems from a limitation of the original item pool from which these scales were constructed. Items represented in existing attachment inventories apparently do not assess security with the same degree of fidelity as insecurity. An important next step for future research on scale development is to write items that tap the low ends of the Anxiety and Avoidance dimensions with better precision.

⁵ By choosing items with the highest discrimination values, we are inevitably capitalizing on estimation error. It is likely that the information curves for these items would be somewhat smaller if the items were recalibrated in a new sample.

Table 2

Item Response Theory Item Parameter Estimates for the 18-Item Experiences in Close Relationships Questionnaire—Revised (ECR-R) Subscale of Anxiety

Item	Anxiety items	Item parameter estimates						
		α	β_1	β_2	β_3	β_4	β_5	β_6
168	I'm afraid that I will lose my partner's love.	2.79	-1.12	-0.39	0.00	0.45	1.08	1.70
57	I often worry that my partner will not want to stay with me.	2.33	-1.38	-0.52	-0.12	0.41	1.15	1.85
1	I often worry that my partner doesn't really love me.	2.21	-1.07	-0.21	0.25	0.82	1.53	2.11
83	I worry that romantic partners won't care about me as much as I care about them.	2.10	-1.64	-0.76	-0.39	0.19	0.93	1.80
110	I often wish that my partner's feelings for me were as strong as my feelings for him or her.	1.98	-1.32	-0.57	-0.29	0.28	0.86	1.58
245	I worry a lot about my relationships.	1.93	-1.71	-0.73	-0.20	0.36	1.00	1.75
226	When my partner is out of sight, I worry that he or she might become interested in someone else.	1.87	-1.36	-0.45	0.04	0.50	1.32	2.05
142	When I show my feelings for romantic partners, I'm afraid they will not feel the same about me.	1.74	-1.85	-0.89	-0.40	0.16	0.90	1.80
191	I rarely worry about my partner leaving me.	1.50	-1.86	-0.67	-0.06	0.60	1.29	2.26
208	My romantic partner makes me doubt myself.	1.49	-0.72	0.35	0.87	1.68	2.43	3.62
82	I do not often worry about being abandoned.	1.36	-1.69	-0.45	0.11	0.73	1.42	2.22
74	I find that my partner(s) don't want to get as close as I would like.	1.36	-1.38	-0.29	0.16	1.07	1.96	2.99
112	Sometimes romantic partners change their feelings about me for no apparent reason.	1.35	-1.31	-0.18	0.30	1.02	1.73	2.59
89	My desire to be very close sometimes scares people away.	1.35	-0.90	0.11	0.50	1.09	1.91	2.81
78	I'm afraid that once a romantic partner gets to know me, he or she won't like who I really am.	1.34	-0.97	0.10	0.52	1.00	1.65	2.61
99	It makes me mad that I don't get the affection and support I need from my partner.	1.32	-1.52	-0.45	0.03	0.79	1.74	2.69
280	I worry that I won't measure up to other people.	1.24	-1.91	-0.71	-0.29	0.33	1.18	2.25
87	My partner only seems to notice me when I'm angry.	1.24	-0.45	0.83	1.40	2.17	2.86	3.53

Note. Items are sorted by their discrimination (α) values.

Table 3

Item Response Theory Item Parameter Estimates for the 18-Item Experiences in Close Relationships Questionnaire—Revised (ECR-R) Attachment Subscale of Avoidance

Item	Avoidance items	Item parameter estimates						
		α	β_1	β_2	β_3	β_4	β_5	β_6
199	I prefer not to show a partner how I feel deep down.	2.28	-1.22	-0.35	0.06	0.57	1.09	1.84
131	I feel comfortable sharing my private thoughts and feelings with my partner.	2.17	-0.84	0.07	0.63	1.15	1.66	2.28
59	I find it difficult to allow myself to depend on romantic partners.	2.08	-1.76	-0.73	-0.19	0.32	0.97	1.92
265	I am very comfortable being close to romantic partners.	2.03	-1.30	-0.26	0.43	1.06	1.64	2.58
171	I don't feel comfortable opening up to romantic partners.	2.00	-1.32	-0.32	0.26	0.79	1.41	2.32
267	I prefer not to be too close to romantic partners.	1.95	-1.33	-0.21	0.44	1.12	1.64	2.48
201	I get uncomfortable when a romantic partner wants to be very close.	1.94	-1.19	-0.31	0.20	0.74	1.25	2.01
36	I find it relatively easy to get close to my partner.	1.93	-1.23	-0.26	0.42	0.97	1.56	2.44
279	It's not difficult for me to get close to my partner.	1.89	-1.42	-0.37	0.22	0.76	1.34	2.20
119	I usually discuss my problems and concerns with my partner.	1.88	-1.07	-0.04	0.81	1.44	2.15	3.16
238	It helps to turn to my romantic partner in times of need.	1.86	-1.19	-0.08	0.84	1.60	2.22	3.04
14	I tell my partner just about everything.	1.85	-1.05	-0.12	0.45	1.01	1.62	2.48
294	I talk things over with my partner.	1.84	-0.89	0.09	0.84	1.51	2.11	2.83
105	I am nervous when partners get too close to me.	1.84	-1.34	-0.36	0.14	0.68	1.34	2.28
242	I feel comfortable depending on romantic partners.	1.74	-2.06	-1.01	-0.11	0.57	1.21	2.12
220	I find it easy to depend on romantic partners.	1.65	-2.18	-1.05	-0.20	0.55	1.17	2.05
300	It's easy for me to be affectionate with my partner.	1.63	-0.95	0.05	0.61	1.20	1.91	2.89
228	My partner really understands me and my needs.	1.60	-1.77	-0.71	0.26	1.16	1.90	2.86

Note. Items are sorted by their discrimination (α) values.

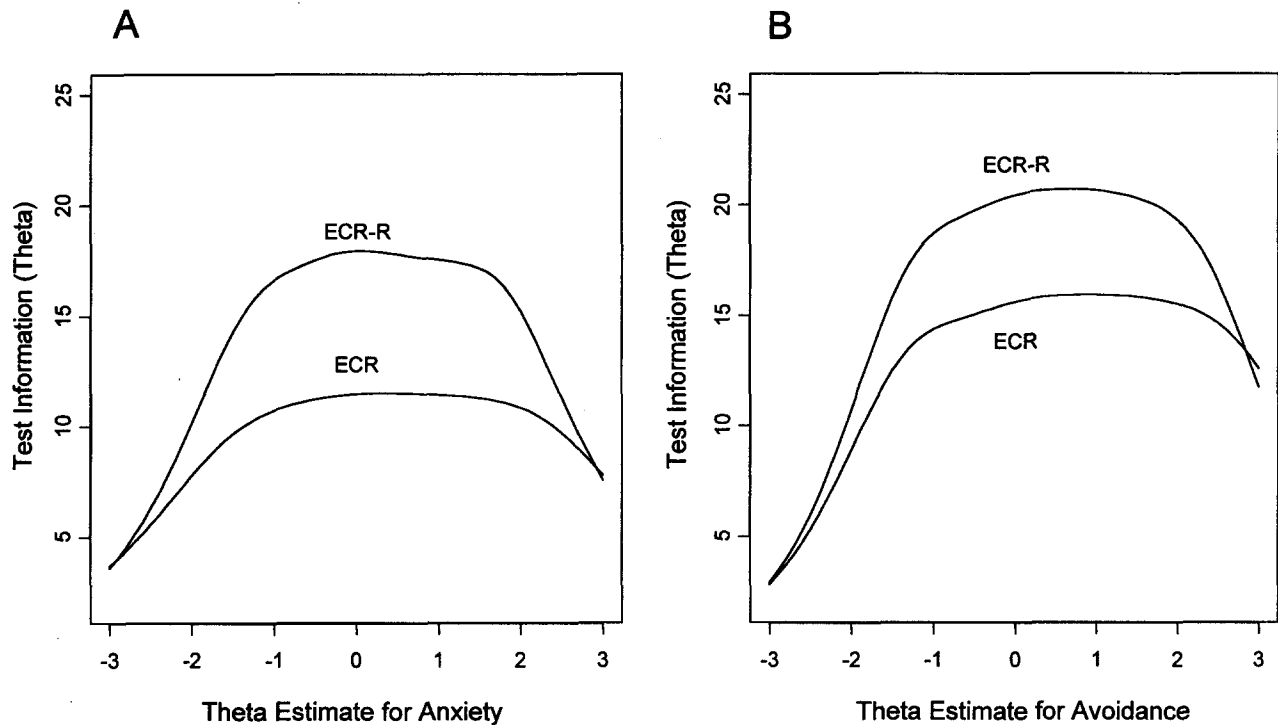


Figure 6. Test information curves for the ECR-R Anxiety and Avoidance scales and the original ECR Anxiety and Avoidance scales.

A second limitation of the ECR-R is that many of the items are conceptually redundant, despite our attempt to prune obviously redundant items from the item pool. Although we believe that there are benefits to probing at certain traits repeatedly to obtain maximally precise measurements, it is desirable to do so by focusing on diverse manifestations of those traits rather than highly specific manifestations of those traits. Again, solving this problem will require the construction of items that are more diverse than those in the current item pool. In the meantime, investigators can easily modify the ECR-R scales by removing what they believe to be undesirable or redundant items. The information properties for modified scales can be assessed easily by plugging the item parameter estimates in Tables 2 and 3 into the equations provided by Samejima (1969, p. 39).

Discussion

Our primary objective in this article was to determine whether existing multi-item self-report measures of adult attachment have the kinds of properties necessary for investigating the theoretical issues typically addressed in attachment research. As we have shown, three of four widely used inventories exhibit undesirable features from an IRT perspective. Specifically, they have a relatively low degree of measurement precision, and, in some cases, they do a poor job of representing the trait continuum with equivalent levels of fidelity. Of the four inventories that we examined, the ECR scales had the best psychometric properties. Nevertheless, we found that the ECR could be improved by using IRT to select

items with optimal psychometric properties. By doing so, we were able to create scales that increased measurement precision by 50% to 100%—without increasing the total number of items.

Do Existing Multi-Item Attachment Inventories Possess Psychometric Properties That Obscure the Interpretation of Data?

We began this article by discussing several ways in which scaling techniques based on traditional methods can produce artifacts that obscure the interpretation of data. It seems appropriate to readdress these issues in light of what we have learned from our IRT analyses. Because these scales contained less measurement precision for highly secure people (i.e., the information functions were not uniform across the trait range), estimates of continuity and differential trait stability may be adversely affected.

To evaluate this possibility, we incorporated the item parameter estimates derived from our IRT analyses of the attachment scales into a series of simulations similar to those discussed previously. (See the section titled Advantages of IRT Over Classical Scaling Methods.) To examine effects of these parameters on test-retest stability, we simulated item responses for each attachment subscale for two time points. For simplicity, the correlation between Time 1 and Time 2 latent trait levels was set to 1.00 (i.e., perfect stability). Item responses for 200 people were generated for two kinds of scales. The first kind of scale used items with parameters identical to those estimated previously in this article. For example,

Table 4
Simulation Results Concerning the Differential Stability of Attachment
for Latent and Observed Scores

Scale and subscale	Test-retest correlation		Differential stability correlation	
	Scale based on well-distributed item parameters	Scale based on actual item parameters	Correlation between change in latent trait levels and Time 1 latent trait levels	Correlation between change in observed scores and Time 1 latent trait levels
ECR				
Anxiety	.91	.94	-.58	-.13
Avoidance	.90	.91	-.58	-.18
AAS				
Depend	.80	.82	.58	.29
Anxiety	.68	.76	-.58	-.02
Close	.78	.82	.58	.32
RSQ				
Secure	.47	.58	.58	.16
Fearful	.77	.78	-.58	.00
Preoccupied	.44	.56	-.58	-.06
Dismissing	.57	.66	.58	.17
Simpson inventory				
Secure	.63	.70	.58	.25
Avoidant	.73	.76	-.58	.06
Anxious	.75	.75	-.58	.05
ECR-R				
Anxiety	.93	.94	-.58	-.23
Avoidance	.95	.95	-.58	-.19

Note. The results for each scale are averages from 100 simulations. For each simulation, latent trait values were keyed such that highly insecure people were more stable than less anxious people. ECR = Experiences in Close Relationships Questionnaire; AAS = Adult Attachment Scale; RSQ = Relationship Styles Questionnaire; ECR-R = Experiences in Close Relationships Questionnaire—Revised.

we simulated item responses to the AAS Close scale by using the item parameter estimates obtained for that scale in our prior analyses. The second kind of scale used items with the same discrimination values, but with well-distributed difficulty values (the difficulty values within an item were evenly spaced between -3.00 and 3.00). The results of these simulations are summarized in Table 4. Notice that for almost every attachment scale, test-retest estimates of continuity are higher for scales based on the estimated difficulty values than scales based on evenly spaced difficulty values. In other words, the relative imprecision of measurement for highly secure individuals is sufficient to artificially inflate the observed degree of continuity. It is noteworthy that the subscales of the ECR and the ECR-R are the least susceptible to this problem.

Davila et al. (1997) observed that highly anxious individuals were less stable in their attachment patterns over time than people who were not anxious. As we suggested previously, this finding could be the result of differential measurement precision across the latent continuum of anxiety. To determine whether the degree of differential measurement precision present in existing attachment scales poses problems for studying differential stability, we conducted a simulation similar to the one we discussed in the section Advantages of IRT Over Classical Scaling Methods. Specifically, we generated latent trait values representing attachment security for 200 people across two time points. People with latent security levels greater than zero at Time 1 did not change at Time 2 (i.e., they were perfectly stable). For people with trait levels less than

zero at Time 1, Time 2 trait levels were constructed to correlate .50 with their Time 1 trait levels. Thus, people high on the trait exhibited perfect stability, whereas people low on the trait exhibited considerable change. (We reversed this pattern for scales keyed in the "secure" direction. For example, for the latent trait of Secure measured by Simpson's scale, people with high latent trait levels exhibited considerable change, whereas people with low secure levels were perfectly stable.)

Item responses for each scale were generated at each time point using the item parameters estimated previously (see Tables 2 and 3). We created an index of differential stability in the observed scores by correlating the absolute difference between Time 1 and Time 2 total scores with Time 1 latent trait levels. Thus, positive correlations indicate that people high on the latent trait at Time 1 exhibited more absolute change in their observed scores from Time 1 and Time 2 than people low on the latent trait at Time 1. Negative correlations indicate that people high on the latent trait at Time 1 exhibited more stability in their observed scores.

As can be seen in Table 4, most of the attachment scales we examined accurately revealed that people high in security were more stable than insecure people. However, the ability of these scales to detect this pattern was limited, especially for the shorter scales (e.g., AAS). Simpson's Avoidant and Anxious scales actually exhibited a pattern of differential stability opposite to that observed at the latent trait level. On the basis of these observations, we suggest that empirical evidence for the differential instability of anxious attachment (Davila et al., 1997) should be reconsidered. If

differential stability in anxious attachment exists, previously used scales are not able to detect it unambiguously.⁶

A couple of caveats are in order. First, we believe the ECR-R can be improved in a number of ways. For example, most of the items for measuring latent anxiety are worded in a trait-positive direction. Future research should focus on developing items that are worded in the trait-opposite direction (i.e., reverse keyed). Also, future research should aim to develop more discriminating items in the secure region of the two-dimensional space. More items are needed to measure the low ends of the Anxiety and Avoidance dimensions.

Second, although we have emphasized the advantages of IRT, it is worth noting that there are several practical and theoretical limitations to IRT that do not apply to classical test theory (Hambleton & Jones, 1993). Unlike classical test theory, many IRT models assume that the construct being measured is unidimensional. Although measurement efforts in personality typically focus on assessing theoretically distinct dimensions, several personality constructs are inherently multidimensional (e.g., self monitoring). Second, because IRT is a model-based approach to assessment, goodness-of-fit analyses are necessary to ensure that the model provides an adequate fit to the data. Such analyses are not necessary within a classical test theory framework. Finally, mathematical analyses of item characteristics are more tractable within a classical test theory framework, and mathematical tools and software for such analyses are easily accessible.

Attachment researchers and personality psychologists more generally seek to understand the nature of social and personality development. However, as we have shown in this article, it is difficult to characterize the developmental properties of latent variables accurately without specifying the response properties of those variables. Classical test methods can yield scales with undesirable response properties that, in turn, can generate misleading inferences concerning individual differences in trait stability and change. We hope that our analyses convince others of the limitations of classical test theory for investigating theoretical issues in the field of adult attachment and the broader fields of personality and social psychology.

⁶ It should be noted that Davila et al. (1997) used a single-item rating of anxious attachment rather than a multi-item scale in their research.

References

- Baldwin, M. W., & Fehr, B. (1995). On the instability of attachment style ratings. *Personal Relationships*, 2, 247-261.
- Bartholomew, K., & Horowitz, L. M. (1991). Attachment styles among young adults: A test of a four-category model. *Journal of Personality and Social Psychology*, 61, 226-244.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-472). Reading, MA: Addison-Wesley.
- Brennan, K. A., Clark, C. L., & Shaver, P. R. (1998). Self-report measurement of adult attachment: An integrative overview. In J. A. Simpson & W. S. Rholes (Eds.), *Attachment theory and close relationships* (pp. 46-76). New York: Guilford Press.
- Collins, N. L., & Read, S. J. (1990). Adult attachment, working models, and relationship quality in dating couples. *Journal of Personality and Social Psychology*, 58, 644-663.
- Crocker, L., & Algina, J. (1991). *Introduction to classical and modern test theory*. Orlando, FL: Harcourt Brace Jovanovich.
- Crowell, J., Fraley, R. C., & Shaver, P. R. (1999). Measurement of individual differences in adolescent and adult attachment. In J. Cassidy & P. R. Shaver (Eds.), *Handbook of attachment: Theory, research, and clinical applications* (pp. 434-465). New York: Guilford Press.
- Davila, J., Burge, D., & Hammen, C. (1997). Why does attachment style change? *Journal of Personality and Social Psychology*, 73, 826-838.
- Dragow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement*, 19, 143-165.
- Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement*, 20, 201-212.
- Embretson, S. E., & Hershberger, S. (Eds.) (1999). *The new rules of measurement: What every psychologist and educator should know*. Mahwah, NJ: Erlbaum.
- Embretson, S. E., & Reise, S. P. (in press). *Item response theory for psychologists*. Hillsdale, NJ: Erlbaum.
- Fraley, R. C. (1999). *Attachment continuity from infancy to adulthood: Meta-analysis and dynamic modeling of developmental mechanisms*. Manuscript submitted for publication.
- Fraley, R. C., & Shaver, P. R. (1998). Airport separations: A naturalistic study of adult attachment dynamics in separating couples. *Journal of Personality and Social Psychology*, 75, 1198-1212.
- Fraley, R. C., & Shaver, P. R. (in press). Adult romantic attachment: Theoretical developments, emerging controversies, and unanswered questions. *Review of General Psychology*.
- Fraley, R. C., & Waller, N. G. (1998). Adult attachment patterns: A test of the typological model. In J. A. Simpson & W. S. Rholes (Eds.), *Attachment theory and close relationships* (pp. 77-114). New York: Guilford Press.
- Gray-Little, B., Williams, V. S. L., & Hancock, T. D. (1997). An item response theory analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, 23, 443-451.
- Griffin, D. W., & Bartholomew, K. (1994). The metaphysics of measurement: The case of adult attachment. In K. Bartholomew & D. Perlman (Eds.), *Advances in personal relationships: Vol. 5. Attachment processes in adulthood* (pp. 17-52). London: Jessica Kingsley.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practices*, 12, 38-47.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer Academic.
- Hambleton, R. K., Swaminathan, H., & Rogers, J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Hartigan, J. A., & Wong, M. A. (1979). A k-means clustering algorithm. *Applied Statistics*, 28, 100-108.
- Hazan, C., & Shaver, P. R. (1987). Romantic love conceptualized as an attachment process. *Journal of Personality and Social Psychology*, 59, 511-524.
- Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data*. New York: Wiley.
- Kim, Y., & Pilkonis, P. A. (1999). Selecting the most informative items in the IIP scales for personality disorders: An application of item response theory. *Journal of Personality Disorders*, 13, 157-174.
- Klohn, E. C., & Bera, S. (1998). Behavioral and experiential patterns of avoidantly and securely attached women across adulthood: A 31-year longitudinal study. *Journal of Personality and Social Psychology*, 74, 211-223.

- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Main, M., & Goldwyn, R. (1994). *Adult attachment rating and classification system: Manual in draft* (Version 6.0). Unpublished manuscript, University of California at Berkeley.
- MathSoft (1999). *S-Plus 2000 User's Guide*. Seattle, WA: Data Analysis Products Division, MathSoft.
- Meehl, P. M. (1995). Bootstraps taxometrics: Solving the classification problem in psychopathology. *American Psychologist*, *50*, 266–275.
- Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement*, *14*, 45–58.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph Supplement*, *17* (4, Pt. 2).
- Samejima, F. (1996). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer.
- Scharfe, E., & Bartholomew, K. (1994). Reliability and stability of adult attachment patterns. *Personal Relationships*, *9*, 51–64.
- Simpson, J. (1990). The influence of attachment styles on romantic relationships. *Journal of Personality and Social Psychology*, *59*, 971–980.
- Simpson, J. A., Rholes, W. S., & Nelligan, J. S. (1992). Support-seeking and support-giving within couple members in an anxiety-provoking situation: The role of attachment styles. *Journal of Personality and Social Psychology*, *62*, 434–446.
- Steinberg, L., & Thissen, D. (1995). Item response theory in personality research. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* (pp. 161–181). Hillsdale, NJ: Erlbaum.
- Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393–408.
- Thissen, D. (1991). *Multilog user's guide: Multiple, categorical item and test scoring using item response theory* (Version 6.0) [Computer software]. Chicago: Scientific Software.
- Traub, R. E., & Lam, R. (1985). Latent structure and item sampling models for testing. *Annual Review of Psychology*, *36*, 19–48.
- van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer.
- van IJzendoorn, M. H. (1995). Adult attachment representations, parental responsiveness, and infant attachment: A meta-analysis on the predictive validity of the adult attachment interview. *Psychological Bulletin*, *117*, 387–403.
- Waller, N. G., & Meehl, P. E. (1998). *Multivariate taxometric procedures: Distinguishing types from continua*. Thousand Oaks, CA: Sage.
- Waller, N. G., & Reise, S. P. (1989). Computerized adaptive personality assessment: An illustration with the absorption scale. *Journal of Personality and Social Psychology*, *57*, 1051–1058.
- Waller, N. G., Tellegen, A., McDonald, R. P., & Lykken, D. T. (1996). Exploring nonlinear models in personality assessment: Development and preliminary validation of a negative emotionality scale. *Journal of Personality*, *64*, 545–576.
- Waters, E., Hamilton, C. E., Weinfield, N., & Sroufe, L. A. (in press). The stability of attachment security from infancy to adolescence and early adulthood: General introduction. *Child Development*.
- Wood, J. M., Tataryn, D. J., & Gorsuch, R. L. (1996). Effects of under- and overextraction on principal-axis factor analysis with varimax rotation. *Psychological Methods*, *1*, 354–365.
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.
- Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, *23*, 299–325.
- Zickar, M. J. (1998). Modeling item-level data with item response theory. *Current Directions in Psychology*, *7*, 104–109.

Received April 7, 1999

Revision received June 24, 1999

Accepted June 24, 1999 ■

Low Publication Prices for APA Members and Affiliates

Keeping you up-to-date. All APA Fellows, Members, Associates, and Student Affiliates receive—as part of their annual dues—subscriptions to the *American Psychologist* and *APA Monitor*. High School Teacher and International Affiliates receive subscriptions to the *APA Monitor*, and they may subscribe to the *American Psychologist* at a significantly reduced rate. In addition, all Members and Student Affiliates are eligible for savings of up to 60% (plus a journal credit) on all other APA journals, as well as significant discounts on subscriptions from cooperating societies and publishers (e.g., the American Association for Counseling and Development, Academic Press, and Human Sciences Press).

Essential resources. APA members and affiliates receive special rates for purchases of APA books, including the *Publication Manual of the American Psychological Association*, and on dozens of new topical books each year.

Other benefits of membership. Membership in APA also provides eligibility for competitive insurance plans, continuing education programs, reduced APA convention fees, and specialty divisions.

More information. Write to American Psychological Association, Membership Services, 750 First Street, NE, Washington, DC 20002-4242.