Elliott, R., Fox, C.M., Beltyukova, S.A., Stone, G.E., Gunderson, J., & Zhang, Xi. (2006). Deconstructing therapy outcome measurement with Rasch analysis: The SCL-90-R. *Psychological Assessment*, *18*, 359-372. DOI: 10.1037/1040-3590.18.4.359

Deconstructing Therapy Outcome Measurement with Rasch Analysis of a Measure of General Clinical Distress: The SCL-90-R

Robert Elliott
University of Strathclyde, Glasgow, Scotland
Christine M. Fox, Svetlana A. Beltyukova, Gregory E. Stone, Jennifer Gunderson, & Xi Zhang
University of Toledo

Author note. We thank the members of the two research teams who helped collect the data analyzed in this article, as well as Melissa Klein and Emily Breighner for assistance with item severity ratings and Sadie Chatmon for data entry. Address correspondence to Robert Elliott, Counselling Unit, University of Strathclyde, 76 Southbrae Drive, Glasgow G13 1PP, UK. Email: fac0029@gmail.com .

Abstract

Rasch analysis was used to illustrate the usefulness of item-level analyses for evaluating a common therapy outcome measure of general clinical distress, the Symptom Checklist-90-Revised (SCL-90-R, Derogatis, 1994). Using complementary therapy research samples, we found that the instrument's 5-point rating scale exceeded clients' ability to make reliable discriminations and could be improved by collapsing it into a 3-point version (combining scales points 1 with 2 and 3 with 4). This, plus removing three misfitting items, increased person separation from 4.90 to 5.07 and item separation from 7.76 to 8.52 (resulting in alphas of .96 and .99 respectively). Some SCL-90-R subscales had low internal-consistency reliabilities; SCL-90-R items can be used to define one factor of general clinical distress that is generally stable across both samples, with two small residual factors.

Keywords: Measurement, Rasch analysis, clinical distress, therapy outcome

Running head: Rasch Analysis of Clinical Distress

Deconstructing Therapy Outcome Measurement with Rasch Analysis of a Measure of General Clinical Distress: The SCL-90-R

The Symptom Checklist 90 Revised (SCL-90-R; Derogatis, 1975, 1994) is one of the most commonly used psychological assessment instruments today. Psychotherapists use the SCL-90-R to aid in diagnosis, inform treatment planning, and measure treatment outcomes (Derogatis & Savitz, 1999). The checklist has been administered to clients meeting criteria for various diagnoses, with different demographic backgrounds, and in inpatient and outpatient settings (Derogatis, 1994). Traditional methods for assessing the properties of psychological instruments have demonstrated that the SCL-90-R is both reliable and valid for measuring treatment outcomes, resulting in its widespread use by clinicians and researchers alike.

However, although useful, the traditional approaches to development and evaluation of measures are not without flaws. For example, Bond and Fox (2001) offer several critiques of instruments developed using traditional methods. Specifically, it can be argued that the symptom or problem index scores that are used as measures are sample and instrument dependent. Overall, it can be argued that in their current form, these scores are closer to raw or ordinal data in the form of observations or counts than they are to measures in the precise sense of the word (that is, containing "objective abstractions of equal units"; Bond & Fox, 2001, p. 2).

Instrument Validation with Rasch Analysis. In order to address some of these shortcomings and to provide a viable alternative to traditional methods, Danish mathematician Georg Rasch (1960, 1980) developed a new model, for measurement in the social sciences, known today as a family of Rasch models and as the basic form of Item Response Theory (IRT). The family of Rasch models provides a framework within which test developers can assess the utility of their measures. The underlying theory of most Rasch models specifies that useful measurement consists of a unidimensional construct arranged in a monotonically increasing pattern (e.g., more than/less than) along an equal-interval continuum (although some multidimensional Rasch models have recently been introduced). If the data fit the Rasch model, they can then be interpreted in terms of abstract, equal-interval units by log transformations of raw data odds and probabilistic equations (within standard error estimates) (Bond & Fox, 2001).

Rasch modeling is a complex topic and the subject of a large and growing scientific literature (e.g., Bond & Fox, 2001; Fischer & Molenaar, 1995; E.V. Smith & R.M. Smith, 2004; Wright & Stone, 2004). Briefly, instruments calibrated using Rasch modeling enable us to determine the extent to which the items have consistently measured a single variable from easy to difficult in a monotonically increasing fashion. Although debates continue, advocates of Rasch modeling argue that this cannot be achieved either with Classical Test Theory or with two or three parameter IRT models. Classical Test Theory is insufficient because it treats ordinal-level data as if they were interval-level, while merely assuming (rather than empirically testing) that the construct possesses an additive structure (Michell, 1997). According to this view, IRT models are not suitable either for constructing measures because they allow for different discriminations among items in the model, thus ignoring a fundamental requirement of measurement (Anderson, 1977). Thus, the two and three parameter IRT models have demonstrated the production of stable measures only when the added parameters (guessing and discrimination) are held arbitrarily constant (Shaw, 1991). Although two and three IRT models can be useful for describing data structure, the family of Rasch models are currently the only available tools for constructing additive scales and diagnosing the extent to which our data fit the fundamental conception of measurement (Bond & Fox, 2001).

Instruments like the SCL-90-R typically use 5-point rating scales as empirical vehicles through which the client can express the nature of experiences relevant to the assessed construct. If we are to trust in the data provided, we must be reasonably assured that clients have made use of the rating scale in the manner intended by the scale developers. However, assumptions that clients make use of the scale as the developers intend are often unsubstantiated. Semantic differences of interpretation (e.g., the specific meaning of the term "moderately" varying with each client) are routinely observed. More egregious are assumptions that clients are able to distinctly and consistently differentiate between rating scale categories (e.g. *moderately* and *quite a bit*). Client confusion when interacting with rating scales has been well documented in the literature (Low, 1988). It is therefore important to investigate whether clients can provide information through the chosen vehicle before using data from an instrument such as the SCL-90-R.

Beyond the rating scale, Rasch statistics assist in the evaluation of the constructed metric; that is, they allow researchers to evaluate the extent to which the items in the measure function unidimensionally. Rasch *fit statistics*, for example, determine whether each item meaningfully contributes to the measurement of a single construct by assessing the extent to which an item or person performs as expected. With adequate fit, easy items are endorsed by more people than are difficult items. Likewise, respondents with more of the measured construct (e.g., psychopathology) endorse more of the "difficult" items (more severe symptoms/more distress) than respondents with less of the measured construct.

Rasch *reliability* estimates include different ways of representing reliability by using either *reliability coefficients, separation (G),* or *number of strata* (E.V. Smith, 2001). The item and person reliability indices estimate the replicability of item placements and person ordering to the universe of similar persons and items. The separation (G) and strata indices estimate the ability of the items to assess different levels of the measure on a less-to-more continuum, and identify the number of subgroups of persons and items that the instrument can discriminate. All of these indices are transformations of one another and help to describe reliability in slightly different ways (E.V. Smith, 2001). Furthermore, unlike their traditional counterparts, these Rasch reliabilities, along with Rasch estimates of item difficulty and person ability, are based on linear measures rather than raw or ordinal data and therefore are more suitable for parametric calculations of means and *SD* (Merbitz, Morris, & Grip, 1989).

Finally, Rasch analysis can identify gaps in the construct continuum by identifying items and persons that are not well targeted. An item is said to be *targeted* when there is a sufficient number of persons at an ability level comparable to the item's difficulty such that the item's difficulty can be more accurately estimated. A person is said to be targeted when there are items with difficulties comparable to the person's ability level. Where items and persons are not well targeted, they have larger than desirable error estimates (i.e., errors associated with different levels of the measured continuum), which indicates gaps in the instrument item set or sample. These gaps provide feedback on how well the instrument is actually measuring what it is supposed to measure within given ranges of the measure, and also what might be done to further improve it. Thus, Rasch analyses provide a useful framework for assessing many of the evidential aspects of validity delineated in Messick's (1995) unified theory of validity (See Bond & Fox, 2001; E.V. Smith, 2001, and Wright & Stone, 2004.)

For these reasons, applying Rasch analysis to psychotherapy outcome instruments such as the SCL-90-R can be beneficial for both researchers and clinicians. Extensive psychometric research on the SCL-90-R using traditional methods of evaluation (see Method section) has

demonstrated that the SCL-90-R yields reliable and valid scores that are sensitive to measuring change in therapy, but questions the discriminant validity of the subscales. In addition, to our knowledge, the SCL-90-R has never been subjected to Rasch or other Item Response Theory analyses. Our general purpose was therefore to develop a detailed and useful understanding of the SCL-90-R by applying item level and conceptual analyses made possible with Rasch analysis. In the present study, we used several forms of Rasch analysis to enhance our understanding of the strengths and limitations of the SCL-90-R, and to illustrate more generally the utility of this approach for psychotherapy outcome measurement. Specifically, the following questions were addressed:

1. Rating Scale Points. What is the optimal number of rating scale categories for the SCL 90-R?
2. Internal Reliability. Can we improve the internal reliability of the SCL-90-R (and its subscales) by dropping misfitting items and unnecessary scale points?
3. Separation/Range. How many distinct clinical groups (strata) can be distinguished using the SCL-90-R?
4. Measurement gaps. What measurement gaps and redundancies exist along the SCL-90-R distress continuum (and those of its subscales), indicating the need for adding or deleting certain types of items?
5. Sampling gaps. For a given sample, what sampling gaps (and redundancies) exist along the SCL-90-R distress continuum (and those of its subscales)?
6. Construct Validity/Theory development. Given the absence of an explicit guiding theoretical model, what monotonically increasing structure (with implications for the sequence of change in therapy) can be suggested for the SCL-90-R (and key subscales such as Depression) using Rasch analysis?

Method

Participants and Procedure

We used clinical samples from two different psychotherapy outcome studies (Elliott et al., 1990; Elliott et al, 2002), both conducted at The University of Toledo.

Depression Sample. Forty-eight clients were primarily recruited through advertisements in local newspapers. Eleven of the clients were men, and 37 were women; their mean age was 36.2 years (*SD* = 11.1); three were Hispanic American, one was African-American and the rest were European-American. All of the clients either fit the diagnostic criteria for current major depressive disorder or were diagnosed with related affective disorders, that is, minor depression (Diagnostic and Statistical Manual of Mental Disorders, 1987) or atypical bipolar disorder (i.e., current major depressive episode plus a history of hypomanic symptoms). Clients were excluded for a variety of reasons (e.g., previous psychiatric hospitalization or active suicidal state). The resulting sample consisted primarily of moderately distressed clients (pretreatment SCL-90-R GSI M = 1.43; *SD* = .45; score range =.41 – 2.19).

Clients completed the SCL-90-R and several other outcome measures twice before beginning therapy, halfway through treatment (after session 8), at the end of treatment (usually session 16), and at 6- and 18-month follow-ups. Because this was a repeated measures design consisting of "stacked samples," clients contributed 1 - 7 SCL-90-Rs (M= 5.0, *SD* = 1.3), for a total n of 139 administrations. (The sample included 28 forms missing 1 to 4 items; one was missing 28 items.)

Naturalistic Sample. The other sample consisted of 72 clients; 62 completed a demographic questionnaire: 36 of these were women; mean age was 43.3 years (*SD* = 13.3); one was Hispanic American, five were African American and the rest were European American; 28 listed a current medication for a psychological condition. Admission criteria were liberal and clients were seen for a variety of Axis I and Axis II disorders. The most common diagnoses were affective (84%) or anxiety (53%) disorders; 44% had Axis II disorders (multiple diagnoses were common). A very small number of clients were excluded because they were actively suicidal, already receiving counseling services, or were diagnosed with acute primary substance or alcohol dependence. This sample thus consisted of a wide range of clinical distress, from apparently nondistressed to severe (SCL-90-R pretreatment M = 1.13; *SD* = .74; score range = 0 – 3.39).

Prior to beginning therapy, and after each block of 10 sessions, clients completed several self-report outcome measures, including the SCL-90-R. Treatment outcome was assessed every 10 sessions via self-report measures. Clients received from 1 to 50 sessions (M = 14.2 sessions). Clients contributed from 1 and 7 SCL-90-Rs (M= 2.8, *SD* = 1.8), for a total n of 159 administrations. (The sample included 26 forms missing 1 to 4 items.)

Instrument

The SCL-90-R (Derogatis, 1994) consists of 90 items rated on a 5-point adverb-anchored rating scale (see Table 1), ranging from (0) *not at all* distressed to (4) *extremely* distressed, for symptoms experienced over the past week. The SCL-90-R includes items similar to "Isolated" (Depression) and "Thinking that others are unreliable" (Paranoid Ideation). (For copyright reasons, these and the items presented in Table 5 and Figure 3 are abbreviated paraphrases of their general meaning; for an accurate interpretation of the results, please consult the actual SCL-90-R items, Derogatis, 1994.)

As typically scored, the SCL-90-R is made up of nine symptom subscales and three overall indices (only one of which is commonly used). Distress is thus treated as a multi-faceted concept evaluated both globally and by breaking it into constituent parts. The overall score for clinical distress, the Global Severity Index (GSI; the mean of all endorsed items) is the most commonly used in therapy outcome research. The subscales are Somatization, Obsessive-Compulsive, Interpersonal Sensitivity, Depression, Anxiety, Hostility, Phobic Anxiety, Paranoid Ideation, and Psychoticism. These are primarily used for diagnostic purposes, but sometimes a particular subscale (e.g., Depression) is used separately in research on a targeted clinical population. Clinical interpretation of the checklist is usually based on a combination of a client's response to the individual items, nine symptom subscales, and the GSI (Derogatis & Savitz, 1999). Thus, from a measurement point of view, the nine subscales collectively represent distress as a domain, and although each may also be measured independently, it is their combined effect that enables the therapist to understand the client's general level of distress. Thus, in actual practice the instrument is typically assumed to be a measure of a single, unitary construct.

The SCL-90-R began as the Hopkins Symptom Checklist (HSCL), developed by Derogatis and colleagues in 1974 (Derogatis, Lipman, Rickels, Uhlenhuth, & Covi, 1974). The original 58-item HSCL was developed to measure the constructs of anxiety, depression, anger-hostility, and obsessive-compulsiveness/phobia (for information on development, reliability, and validity indices, see Derogatis et al., 1974). The SCL-90 was then developed by adding scales for somatization, schizophrenia and paranoid ideation (Derogatis, Lipman, & Covi, 1973) and

was subsequently revised (SCL-90-R; Derogatis, 1977, 1994). The revised version (SCL-90-R) was normed using four different groups of people: psychiatric inpatients and outpatients, community nonpatient adults, and community adolescents (Derogatis, 1994).

Traditional psychometric analyses of the SCL-90-R have consistently reported acceptable levels of internal consistency. For instance, Derogatis, Rickels, and Rock (1976) administered the SCL-90 to 219 symptomatic volunteers and found coefficient alphas in the range of .77 (Psychoticism) to .90 (Depression). A study of 103 community outpatients reported coefficient alphas ranging from .79 (Paranoid Ideation) to .90 (Depression) (Horowitz, Rosenberg, Baer, Ureno, & Villasenor, 1988). Test-retest coefficients have also been calculated at both 1 and 10 weeks. One-week retests with 94 outpatients reported coefficients in the range of .78 (Hostility) to .90 (Phobic Anxiety) (Derogatis, 1994). Ten-week retests with 103 outpatients found coefficients in the range of .68 (Somatization) to .83 (Paranoid Ideation), with the General Severity Index at .84 (Horowitz et al., 1988).

Convergent validation research indicates that the SCL-90-R correlates with many instruments measuring similar constructs. For instance, Boleloucky and Horvath (1974) compared the SCL-90 to the Middlesex Hospital Questionnaire and found good convergence and discrimination between the two scales. Derogatis and colleagues (1976) compared the SCL-90 with the Minnesota Multiphasic Personality Inventory (MMPI) clinical scales, as well as the MMPI content and cluster scales. Convergent validity with the MMPI scales was found for 8 of the 9 SCL-90 symptom subscales among a group of 209 symptomatic volunteers. The only exception was the SCL-90 Obsessive-Compulsive subscale, which does not have a comparable MMPI scale.

Convergent validity research has also focused on specific SCL-90-R indices. In particular, the Depression subscale has been found to correlate with the General Health Questionnaire Depression scale (Koeter, 1992), the Asperg Rating Scale (Peveler & Fairburn, 1990), and the Beck Depression Inventory (Choquette, 1994; Peveler & Fairburn, 1990). The anxiety scales of the SCL-90-R and the General Health Questionnaire have also demonstrated correlations (Koeter, 1992). Finally, the global indices were found to correlate with the Present State Examination, a clinician administered structured interview (Peveler & Fairburn, 1990).

One of the major uses of the SCL-90-R is as a mental health outcome measure. There is evidence that the instrument is sensitive to change in both psychotropic medications (e.g., Kim & Dysken, 1990; Levine, Anderson, Bystritsky, & Barton, 1990; Walsh, Hadigan, Devlin, Gladis, & Roose, 1997) and psychotherapy (e.g., Crits-Cristoph, 1992; Kopta, Howard, Lowrey, & Beutler, 1994). Kopta and colleagues (1994) used the SCL-90-R to measure change in 854 outpatient psychotherapy clients. The researchers found that clinical symptoms responded to psychotherapy before life functioning. Specific symptoms that showed early improvement were anxiety, depression, obsessive-compulsiveness, interpersonal problems, and somatization. Symptoms that responded more slowly to psychotherapy included hostility, paranoid ideation, psychoticism, sleep disturbance, and overeating.

However, the SCL-90-R's subscales consistently show large intercorrelations, suggesting that they are not conceptually distinct and casting doubt on their discriminant validity. For example, studies by Dinning & Evans (1977), Holcomb, Adams, and Ponder (1983), and Clark and Friedman (1983) reported mean intercorrelations among subscales or comparable factors ranging from .59 to .67. Furthermore, whereas some factor analytic research has supported the hypothesized subscale structure of the SCL-90-R, other studies have not. For example, when Derogatis and Cleary (1977) used Procrustes and Varimax rotations with 1,002 outpatients, they

were able to extract the expected subscale structure, with the exception of the Psychoticism subscale. However, other researchers have found somewhat different factors, such as a primary factor of overall distress, indicating that the SCL-90-R measures client distress more generally (e.g., Brophy, Norvell, & Kiluk, 1988).

Data Analysis

The data were analyzed using the Rasch rating scale model (Andrich, 1978) in the WINSTEPS (Linacre & Wright, 2004) software. WINSTEPS works around missing data, treating them as nonadministered. For this analysis, the samples were combined. As noted, the SCL-90-R is used to measure change over time, so we needed the sample to include multiple data points for each client, yielding a total n of 298 observations from 120 different clients. Examination of the observed frequencies of each rating category per item suggested that this sample size was sufficient, following recommendations that Rasch analyses should have at least 10 observations for each response category (in this case 5 X 10 = 50; Linacre, 2002). However, because of nonindependence of observations, statistical significance levels and n-based error estimates should be interpreted cautiously. (Using Rasch person fit statistics to probe for possible nonindependence, we identified 29 observations – about 10% of the sample -- with substantial overfit of less than .6, indicating that they were overly-predictable.)

Separation and Reliability. First, we examined Rasch person and item separation statistics, G, for the entire instrument, to determine the level of distinction possible among persons and items along the measured variable; then we analyzed each subscale. Separation is the ratio of the square root of the variance explained by the measurement model ("adjusted person variability") to that of the unexplained variance or measurement error including error from model misfit ("real root mean square error"), that is, the signal-to-noise ratio. Because separation is open-ended, it does not suffer from the ceiling effect problem of alpha-type reliability estimates. In addition, Rasch separation statistics can be transformed into a strata index, which determines the number of statistically different levels of person ability that are distinguished by the items (Strata = [4G +1]/3; Fisher, 1992; E.V. Smith, 2001; Wright & Masters, 1982). A separation of 2.0 (i.e., identifying 3 strata) is considered to be the minimum acceptable value (Wright & Masters, 1982). Person-item maps were also examined to help with interpretation of the person and item separation statistics as well as to understand item ordering, sampling and measurement gaps.

Person and item reliabilities were also considered and expected to be high because of the length of the instrument. As Linacre (1996) has noted, although true score (alpha) reliability and Rasch reliability are estimates of the same coefficient and are interpreted in the same way, a subtle yet important difference in their handling of extreme scores exists: Traditional measurement assumes zero error variance associated with extreme measures, whereas Rasch treats these extremes as missing. Rasch reliability is thus more conservative in this respect.

Rating Scale Category Analyses. Second, in order to evaluate the functioning of the rating scale categories, we used common Rasch rating scale diagnostics to examine how the clients used the 5-point rating scale. The most common diagnostics focus on the category thresholds, that is, the estimated difficulties in choosing one response over another (for example, the difficulty in choosing *strongly agree* over *agree*; Wright & Masters, 1982). Thresholds should increase monotonically (i.e., should be ordered in the same manner as intended by the item developer) and should be appropriately distanced from one another (i.e., should be at least

1.4 logits apart, but not farther than 5 logits, Linacre, 1999) if the items are to measure distinct and meaningful progression along the variable.

Another rating scale diagnostic that we used entails visually detecting useful distinctions among response categories by looking at probability curves (see Figure 1). These show the probability of choosing a given rating scale category for every place along the measured variable. Useful categories are those with high probabilities that span a distinct portion of the variable (Bond & Fox, 2001). Categories that overlap too much with adjacent categories are typically not helpful in defining a distinct point along the variable. Thus, in Figure 1, the horizontal axis is the distress level of the client minus the distress level associated with the item (which we will refer to as adjusted clinical distress). For example, if a client's distress level is 3, 2, or 1 units higher than the distress level measured by a particular item (corresponding to a 3, 2, or 1 on horizontal dimension in Figure 1), the most probable response is to endorse a '4' (corresponding to probabilities ranging from about .5 to .9 on the vertical dimension). Conversely, if a client's distress level is 3, 2, or 1 units *lower* than the distress level measured by an item (corresponding to a –3, –2, or –1), the most probable response is endorse a '0' (corresponding to probabilities ranged from about .45 to .9). Therefore, the probability curves visually display the same information as the table of thresholds.

Item Fit Analyses. To determine if any of the items on SCL-90 captured something qualitatively different from overall distress, infit mean squares were examined, using the value of 1.4 as the cutoff for rating scales (Bond & Fox, 2001). On the other hand, item redundancy was investigated by the outfit statistics using the similar criteria as well as by the largest standardized residual correlations, after partialing out the measured dimension of general clinical distress. Combined with the information of item fit, the largest standardized residual correlations provided guidelines for how to shorten the SCL-90 with the least loss of information. Thus, we used the best fitting item of each pair to identify the items that should be retained (i.e., better fitting items) as well as those that can be used to shorten the instrument, using a criterion of .40 for residual correlations.

Construct Analysis. Additional evidence of construct validity of SCL-90 was obtained by conducting a qualitative analysis of the items ranked by difficulty. This allowed us to see if the obtained ordering of item clusters made clinical and theoretical sense. Finally, since the closest analog of the psychological distress continuum measured by the SCL-90-R appeared to us to be the General Assessment of Functioning (GAF: Axis V of the *Diagnostic and Statistical Manual-IV*; American Psychiatric Association, 1993), we asked three graduate students in clinical psychology to rate each of the SCL-90-R items on the GAF. To do this, we told them to assume that a client presented with the symptom described by the item at a moderate level of severity and then to estimate such a client's GAF (in 5-point increments on the 100-point scale).

Analysis of Unidimensionality of SCL-90. The dimensional structure of the SCL-90 was investigated in two ways. First, Rasch fit statistics and score correlations of each item with the latent variable (e.g., clinical distress) were reviewed as recommended by E.V. Smith (2002). Not only do these measures provide an indicator of consistency across the criterion measured by the instrument, but by extension, they begin to address the possible existence of multiple dimensions. Items that misfit and those with small or even negative score correlations may simply be written poorly, but are often indicative of the presence of another underlying variable.

The second approach was the use of Rasch fit in conjunction with Rasch principal components analysis (RPCA). RPCA can uncover the presence of multiple dimensions, although it cannot alone determine whether a factor is an underlying aspect within the larger

construct or whether the uncovered dimension is a unique construct separate from the main one we intended to assess. Thus, we performed an analysis of response residuals among items in order to see if we could find any evidence for the presence of unsuspected secondary variables, after removing variance due to the primary measured distress dimension (Wright & Stone, 2004). The results of this analysis showed how much variance was explained by the single overall measurement dimension. The analysis of residuals also allowed us to identify if any additional clusters of the items might be present.

Comparability of the Samples. Finally, we concluded our investigation by looking at the comparability of the two samples as a measure of clinical distress. This analysis was necessary because the SCL-90-R is used with different groups of psychotherapy clients. Therefore, "it is essential that the identity of the variable be maintained from one occasion to the next" (Wright & Masters, 1982, p. 114). By testing for statistically significant differences in item estimates obtained from separate Rasch analyses on each sample (referred to as a "DIF analysis"), we assessed whether the items had significantly different meanings for different groups of clients.

Results

Diagnostic Analyses

Separation and Reliability Analyses. Overall, Rasch person and item separation statistics, G (4.90 and 7.76 respectively) showed a high level of distinction among persons and items along the measured variable. The person separation of 4.90 translates into 6 statistically distinct strata, whereas the item separation of 7.76 translates into 10 distinct strata. Person and item reliabilities were also high as expected, corresponding to alpha values of .96 and .98, undoubtedly because of the length of the instrument.

Scale Category Analyses. The response categories followed the expected progression of rated levels, that is, they advanced monotonically from *not at all* to *extremely*, as Table 1 indicates. However, as the step threshold estimates in this table show, the two adjacent categories 2 (*moderately*) and 3 (*quite a bit*) were not statistically significantly different (i.e., they were only .05 logits apart, $t = 1.77$; df = 5620), indicating that the clients did not reliably distinguish between these categories.

Examination of the probability curves (see Figure 1) revealed that categories 1, 2, and 3 are the most probable categories across only a very small section of the variable (from about –.6 to .8), but their highest probability of endorsement reaches only about .35. Categories 2 and 3 are the most redundant visually, thus suggesting the same conclusion we reached by examining the table of thresholds.

Misfit Analyses. Eleven items showed significant overall misfit with the measure. This suggests that these items captured something qualitatively different from overall distress. Indeed, they varied greatly in content (paraphrased as: overindulging on food, sexual issues, auditory hallucinations, fear of public transportation, ill-as-ease consuming food or drink with others, fear of leaving home by oneself, fear of passing out with others, weeping, holding ideas not yours, and feeling others direct your thinking), and they included some of the highest severity items. No items suffered from overfit (a sign of redundancy, which could have artificially increased internal consistency).

Improving the Instrument

To fix the problems with the original 5-point rating scale, we initially combined rating categories 2 (*moderately*) and 3 (*quite a bit*) as the closest to each other. However, this

recategorization was not optimal, using the criteria outlined by Lopez (1996), in that the best discrimination of the rating scale and the best data-model fit were not achieved. Table 2 summarizes the ways of collapsing rating categories that were attempted before settling on a parsimonious 3-point scale, arrived at by combining categories 1 (*a little bit*) and 2 (*moderately*), and also categories 3 (*quite a bit*) and 4 (*extremely*). Combining these categories also made sense conceptually. Repeating the analysis with the proposed 3-point scale revealed that six out of the eleven previously identified items with problems improved their rating scale functioning, but 3 items still misfit. Overall, the strategy of category collapsing into a new 3-point scale and removing the three remaining misfitting items produced an increase in person separation from 4.90 to 5.07 (the latter corresponding to an alpha of .96 and identifying 7.09 strata of clients) and in item separation from 7.76 to 8.52 (corresponding to an alpha of .99 and identifying 11.69 strata of items). In addition, the distinctiveness of each newly formed response category increased, as seen in Table 3 and Figure 2, where each category peaks and is, therefore, the most likely response choice at some part of the measured continuum. Additional research is needed to cross-validate the functionality of this newly formed scale.

Levels of Distinct Client and Item Severity

We next examined the question of interpreting the person and item separation statistics and the item order, as depicted on the person-item map (Figure 3a & 3b). As noted, the revised instrument with a 3-point rating scale can distinguish seven statistically distinct groups or strata of clients when using 90 items of the SCL-90R. This indicates that, as an instrument, the SCL-90-R is probably capable of discriminating among intuitively obvious groups such as nonclinical, mild, moderate, serious, and extreme.

At the same time, the instrument is also capable of discriminating among at least eleven statistically distinct levels or strata of item severity. It was not our intent to provide a precise analysis of these eleven strata; however, to obtain a fuller understanding of the nature of the SCL-90-R's clinical distress construct we used an informal qualitative analysis. To help deal with the large number of items, the first author categorized items with similar difficulty scores by their content, resulting in 11 partially overlapping qualitative clusters of items, ordered roughly by severity. This analysis did not produce discretely ordered categories or strata; instead, at each point along the continuum, there were two or three overlapping item groups (see Figure 3a & 3b), roughly ordered as follows:

1. Psychosis
2. Severe agoraphobia
3. Aggression
4. Serious medical concerns (e.g., feeling too warm/too cool, loss of feeling/prickles, stomach, heart symptoms)
5. Panic/Strong anxiety
6. Moderate anxiety/Major depression
7. Interpersonal problems (externalizing, suspicious, resentful)
8. Mild medical concerns (aches and pains)
9. Moderate depression (cognitive, interpersonal symptoms)
10. Crankiness/irritability
11. Mild depression/General malaise

Nature of Measured Dimension

The obtained rough ordering of item groupings makes clinical and theoretical sense, while at the same time clarifying relationships among disparate clinical symptoms. Using these clusters, the dimension measured by the SCL-90-R can be described qualitatively in terms of roughly four intertwined conceptual strands that encompass the levels of severity just described: (a) an *Anxiety/Depression* strand that ranges from Panic/Strong Anxiety through Mild depression/General malaise, appears to be the longest, and includes 4 item groupings; (b) an *Interpersonal Problems* strand that ranges from Aggression through Crankiness/Irritability and encompasses 3 groupings; (c) a *Medical Concerns* strand that consists of 2 clusters, extending across a wide severity range; and, finally, (d) a *Psychotic/Breakdown* cluster, occurring at the highest severity level, that appears to be the culmination of the other three strands, and to represent prototypical psychological dysfunction, as the opposite pole of the continuum from Mild depression/General malaise. (These qualitative item groupings and strands are presented for descriptive purposes only; we do not see them as structurally coherent units.)

The closest analog of the psychological distress continuum measured by the SCL-90-R appeared to us to be the General Assessment of Functioning (GAF: Axis V of DSM-IV; American Psychiatric Association, 1993). In order to test this assumption and to further examine the SCL-90-R's construct validity, we asked three graduate students in clinical psychology to rate each of the SCL-90-R items on the GAF. To do this, we told them to assume that a client presented with the symptom described by the item at a moderate level of severity and then to estimate such a client's GAF (in 5-point increments on the 100-point scale). The three raters agreed substantially with one another (average measure intraclass correlation, consistency definition = .86). As predicted, the mean of their ratings correlated substantially with the logit severity values for the SCL-90-R items: $r = .63$. This indicates that the SCL-90-R distress severity dimension is strongly related to but not identical to the GAF. Five SCL-90-R items had significant, positive standardized residuals (>2.0), indicating that their Rasch scaled measure values were reliably higher than their GAF-predicted values, paraphrased as: 16. auditory hallucinations; 47. fear of going on public transportation; 53. difficulty swallowing; 73. ill-at-ease consuming food or drink with others; 82. fear of passing out with others. Thus, one finding of this analysis is that GAF ratings may underestimate clients' perceptions of the severity of these symptoms, at least in the estimation graduate student diagnosticians.

Sampling Gaps

Even with the highly screened clinical sample of depressed clients and the wide range of severity of clients in the naturalistic sample, Figure 3 indicates that clients with more severe levels of psychological symptoms and overall distress have been undersampled here. Very few people in the sample had a high probability of endorsing the category *quite a bit* to *extremely* for the items on the SCL-90-R (only about 14% of item responses were in this range). The majority of the scores in the two samples were at the moderate or mild level of distress (see person distribution at bottom of Figure 3).

Measurement Gaps and Redundancies

Although the spread of the items is large (i.e., more than 11 logits), adding some easy items that would capture symptoms of absence of distress would be recommended because some of the clients are at the low end of the scale (low scorers), that is, lower than the items in the SCL-90-R measure (see Figure 3a & 3b). In addition, some items appear to be redundant.

Examination of the largest standardized residual correlations suggested dropping 8 items; most of them very closely related symptoms (e.g., "thinking about causing your own death" vs. "ideas of passing away").

Similarly, a map of the order of all 90 items on the measured dimension (Figure 3) showed that many of the items on the SCL-90 share virtually identical difficulty levels (i.e., the location of "0", "1", and "2" are the same for numerous items), indicating that they are "measure-similar items" (Wright & Stone, 2004). Such items are largely redundant in a measure of the overall psychological distress dimension, although they still contribute statistical information here and may prove useful as part of subscales measuring specific types of psychological distress such as depression.

Evaluation of SCL-90-R Subscales

The analysis of each subscale using the new 3-point rating scale revealed that all subscales contained items with at least 3 degrees of separation (six subscales had item separations of greater than 7.0, i.e., at least 9 strata), indicating that a wide range of item severity was sampled within each subscale. However, person separation was a different matter entirely: six out of nine subscales did not provide a minimum G value of 2.0 (i.e., identifying 3 strata; Wright & Masters, 1982) (see Table 4). The remaining three subscales were barely better than the minimum recommended value: Depression (G=2.29), Interpersonal Sensitivity (G=2.01), and Obsessive Compulsive (G=2.06). In fact, four subscales had inadequate person reliabilities of less than .7 (Hostility, Paranoid Ideation, Phobic Anxiety, Psychoticism); these subscales could distinguish no more than 2 strata of clients.

Evaluation of Unidimensionality and Possible Secondary Variables

The assessment of unidimensionality from the Rasch fit statistics and score correlations of each item with the latent variable showed that only item 60, paraphrased as "overindulging on food", had both a poor fit value (infit >1.4) and a low score correlation < .3. (This was one of the three dropped items.)

The Rasch principal components analysis revealed that the single overall measurement dimension explains 67.7 units (i.e., eigenvalues) of the item variance out of a total of 87 (corresponding to the number of items), i.e., 78% of the total variance. The analysis of residuals also showed that two additional clusters of the items might be present.

The first cluster consisted of seven items with substantial positive loadings (i.e., with off-dimension loading of .4 or greater) and appears to have a common meaning that can be labeled as "Social Distress," paraphrased as: 69. highly ill-at-ease around people; 61. discomfort being observed or discussed; 79. viewing self as without value; 88. always isolated from others; 21. uncomfortable or timid with other gender; 76. accomplishments underappreciated by others; and 37. viewing others as cold or rejecting.

The second cluster included six items with loadings greater than .4 and shared a common meaning, which can be labeled as "Depressive Motivational Deficit," paraphrased as: 71. hard to do anything; 14. sluggish or listless; 54. despairing about what is to come; 30. sadness; 32. did't care about anything; and 55. difficulty focusing thoughts.

These two sets of residual variables thus provide some evidence for two secondary scales within the SCL-90-R and meet the recommended 3-unit criterion (Linacre & Wright, 2004): 5.6 units (6% of total variance) and 4.5 units (5%). However, all the 11 items in the two secondary factors had score correlations greater than .4; six were at least .6. Thus, these items also strongly

measure the overall clinical distress variable. Furthermore, these factors are dwarfed by the overall distress variable, which accounts for almost seven times as much variance as the two combined. These analyses indicate that while the SCL-90-R is not totally unidimensional, its multidimensional components are relatively trivial.

Is the Meaning of Clinical Distress Comparable Across the Samples?

The analysis of the comparability of the two samples revealed that there are eight items that were seen as significantly *more* distressing (higher logit scores; $p < .01$), whereas seven other items were seen as significantly *less* distressing (lower logit scores; $p < .01$) for the naturalistic sample (see Table 5). Both sets of distinctive items are diverse in their content and difficulty levels; no clear pattern of differential meaning is apparent. The differences ranged in size from .5 to .88 logits; none approached the 1.4 logits difference required for a meaningful difference between adjacent scale points, and only seven differences spanned step boundaries at -.94 and .94. Additionally, none of these differences had a large effect size, and only one exceeded .5 (see Table 5). Thus, although there are statistically reliable differences in item meaning across the two samples, those differences amount to less than a scale point (on the revised 3-point scale) and more than 80% of the items did not differ reliably.

Discussion

The purpose of this study was to evaluate the the SCL-90-R using Rasch analysis, as an example of the potential usefulness of applying this approach to measurement to a widely used measure of clinical distress. A PsychInfo search for Rasch research on disordered populations identified 197 studies, most of which dealt with specific symptom instruments (e.g., PTSD, Betemps, R.M. Smith, Baker & Rounds-Kugler, 2003; depression, Cole, Rabin, Smith & Kaufman, 2004). Only four of these involved general clinical distress or symptom instruments; two of these used some form of the SCL-90-R: Olsen, Mortensen & Bech (2004) a Rasch-related forms of homogeneity analysis to look at the SCL-90-R in a nonclinical population; and Mool (1998) used Rasch category analysis with a short form of the SCL-90-R in a small clinical sample. Thus, as far as we have been able to determine, this is the first English-language study to utilize a range of Rasch methods on a broad-band psychological distress instrument using a clinical population.

Rasch analyses provide several unique contributions to understanding the functionality of the SCL-90-R not developed through earlier, traditional analyses. One of the fundamental goals of Rasch analysis is the development of clear, functional, linear variables. Rasch analysis affords the researcher the opportunity to evaluate the clarity of the criterion being measured by the instrument, via both summary statistics (separation and reliability) and the person-item map. Because person and item parameters are estimated as separately from the sample as possible, Rasch estimates are considered relatively "sample-independent" indicators of how well the instrument is able to reflect the desired criterion. Traditional models, by comparison, are not constructed to provide information beyond the sample studied (although this limitation is commonly overlooked). Moreover, Rasch analyses offer the researcher multiple, detailed performance indicators (including various fit and point-biserial statistics) on both individual item and total instrument levels. Use of these resources helps to ensure the functionality of items and may also be used to evaluate the possible existence of secondary variables, via RPCA.

At the same time, it is important to point out that Rasch analysis is just one approach to item response theory (although it is the oldest and simplest approach). In particular, there are

important philosophical differences between Rasch analysis and multiple parameter IRT models. Rasch analysis is a rational-empirical approach to measurement, which specifies that all measurement (physical as well as psychological) is inherently represented by a single parameter (the item difficulty along the construct being measured) and evaluates the extent to which the data fit that parameter. In contrast, multiple-parameter models typically work from a more empiricist stance, seeking to explain the data at hand rather to adhere to a strict definition of measurement, resulting in the development of more complex models of the data (e.g., corrections for guessing). Simply put, Rasch analysis seeks to find a monotonic dimension in the data that can be used for to construct (or improve) a measure; multiple parameter models seek a mathematical model that comprehensively represents the data. Clearly, these are rather different enterprises, both of which have their place. In this article we have focused exclusively on Rasch analysis and its use for improving psychotherapy outcome measurement.

Turning now to our specific findings about the SCL-90-R, the results of the present study support some of the previous research and add new information to our knowledge of the SCL-90-R. Rasch analyses of the rating scale found that the SCL-90-R rating categories advance monotonically from *not at all* to *extremely*. However, we also found that clients did not effectively discriminate between categories 2 (*moderately*) and 3 (*quite a bit*) and that the most effective form of the rating scale collapsed 1 (*a little bit*) with 2 (*moderately*) and 3 (*quite a bit*) with 4 (*extremely*). Ironically, it appears that clients often treated the SCL-90-R more like a checklist (i.e., indicating whether symptoms are present or absent) than as a rating scale.

Person and item separation statistics are unique aspects of Rasch analysis. Qualitative interpretation of the item calibrations identified at least 7 strata categories of person separation, ranging from nonclinical to extreme. This indicates the strength of the SCL-90-R to measure a wide range of clinical distress. These findings corroborate previous research demonstrating that the SCL-90-R measures severity of psychopathology in general (Brophy et. al, 1988). Furthermore, the SCL-90-R was able to differentiate a wide range of clinical populations in spite of undersampling the upper range of client severity. It should be noted that Derogatis' 1994 normative sample suffers from the same problem; and in fact the mean pretreatment score for our depressed sample was higher than the mean of the normative outpatient sample.

Rasch item separation points to the existence of 11 discrete item severity strata, but our qualitative analysis suggested that it might be more accurate to think of these as interwoven moderately overlapping content themes (see Figure 3a & 3b) ranging from mild depression to psychosis. This is one of the benefits of Rasch analysis, clarifying the progressive difficulty of items along the measured dimension. This is also new information, as the SCL-90-R was originally based on diagnostic categories and was therefore not developed as a single-dimension measure, in spite of being typically used in this way. Additionally, the order of item difficulty appears to be somewhat similar to those proposed by Kopta and colleagues (1994) based on SCL-90-R psychotherapy outcome research. Kopta and colleagues found that anxiety, depression, somatic, and interpersonal symptoms change first in therapy, but aggression/hostility and psychotic symptoms respond more slowly. In the current study, Rasch analysis categorized anxiety, depression, somatic, and interpersonal problems as 'easier' items, whereas aggression/hostility, severe anxiety, and psychoticism items were revealed to be more difficult to endorse.

However, difficulty to change is not the same as severity, which is understood in Rasch analysis as difficulty to endorse. Thus, a next step would be to examine the relationship between the Rasch-defined clinical distress dimension found here and sensitivity to change over the

course of therapy. Such a line of further research might have clinical implications if SCL-90-R scores can be used as an indication of recommended treatment dosage (e.g., estimating length of time in therapy or treatment outcome). This would be particularly helpful because, unlike other commonly used clinical instruments, the SCL-90-R does not have a scale that specifically measures receptiveness to treatment.

Furthermore, we found that in addition to serving as a continuum of clinical distress, the SCL-90-R acts in a manner similar to the General Assessment of Functioning (GAF) scale, which measures overall functioning. Further Rasch analyses could be used to create a formula for translating a SCL-90-R general score into an equivalent GAF score or other clinical distress instrument (e.g., Outcome Questionnaire-45; Lambert et al., 1996).

When applying Rasch analysis to the SCL-90-R subscales, we discovered that each subscale had at least 3 degrees of item separation. This indicates that a wide range of item severity is sampled within each subscale. However, person separation or reliability was often not adequate for the subscales, with only depression, interpersonal sensitivity, and obsessive compulsive scales demonstrating adequate (but marginal) person separation. Overall, the subscales are not particularly useful for distinguishing among populations of clients. These current findings complement previous research using traditional methods, which have found intercorrelations among the scales resulting in difficulty discriminating among subscales (e.g., Clark & Friedman, 1983; Dinning & Evans, 1977; Holcomb et. al, 1983). At the same time, the principal components analyses that we ran on the residuals (with the variable for general clinical distress removed) did point to the existence of two relatively small additional subscales, one for a depressive motivational deficit, and the other for social distress. Additional subscales might be constructed using these items to supplement the overall general distress scale in order to provide better differentiation within depressed or interpersonally distressed client populations.

Qualitative analysis of the person-item map revealed four conceptual strands including anxiety/depression, interpersonal problems, medical concerns, and psychotic symptoms. These strands seem to be measuring four important but very distinct areas of clinical concern: emotional, interpersonal, somatic and perceptual/reality testing. That anxiety and depression fell on the same strand is not surprising as recent research supports the similarity and high correlation of these disorders (e.g., Mineka, Watson & Clark, 1998). These four content strands may be more useful than the 8 original subscales. Nevertheless, it appears that item variance is best explained by the overall clinical distress dimension.

The Rasch analyses reported here have added another perspective to previous measure development research on the SCL-90-R, providing a tentative view of the strengths and weaknesses of the instrument. Thus, we recommend that the following possible improvements of the SCL-90-R be explored via further research with a larger and possibly more clinically distressed sample: 1. Collapse rating categories 1 and 2, as well as 3 and 4, to create a 3-point rating scale (i.e., a checklist with two levels of severity). 2. Drop the 3 remaining misfitting items: loss of sexual interest, bothersome ideas about sex, and overeating. (Implementation of these first two points substantially increases person and item separation.) 3. Explore the possibility of dropping 8 redundant items identified as redundant with another item (although such items might prove useful for severely distressed client populations). 4. Add items measuring minimal distress to increase accurate measurement of less distressed clients (perhaps basing items on questionnaires developed for community as opposed to clinical samples, e.g., Center for Epidemiological Studies – Depression scale, Radloff, 2000).5. Develop an algorithm for generating a Rasch-derived Clinical Distress index from SCL90-R raw scores.6. Develop

algorithms for equating SCL-90-R Clinical Distress with the GAF and other measures of general clinical distress. 7. Abandon use of less than adequately differentiated subscales. 8. Develop new subscales for social distress and depressive motivational deficit.

A key limitation of this study is the relative thinness of the sample at the highest severity levels, as is evident in Figure 3a & 3b. Although this problem is shared with the Derogatis' (1994) original normative samples, the consequence is that without sufficient data, the standard errors are larger and the item estimates are not as stable as they should be at the high end of the scale. Our inferences about the item calibrations are more stable at the lower end than at the higher end of the clinical distress dimension. As a result, the positions of higher end items are less likely to replicate. Therefore, further validation of the SCL-90-R would best be targeted at more extreme populations, such as clients presenting at psychiatric crisis centers with suicidality, severe substance abuse or psychotic problems, excluded in the samples analyzed here. These results could then be equated with our findings to construct a more accurate, longer measure for clinical distress.

We suggest that more such analyses be conducted in the future as part of measure development of similar instruments for assessing general clinical distress. Additionally, it would be useful to employ Rasch analyses to other common instruments used to measure clinical disorders, as it appears to be difficult to create subscales that completely differentiate between various diagnoses.

References

American Psychiatric Association. (1993). *Diagnostic and statistical manual of mental disorders* (4th ed.) Washington, DC: Author.

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed. rev.) Washington, DC: Author.

Andersen E.B. (1977). Sufficient statistics and latent trait models. *Psychometrika , 42, 1*, 69-81

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrica, 43,* 357 – 374.

Betemps, E.J., Smith, R.M., Baker, D.G., & Rounds-Kugler, B.A. (2003). Measurement precision of the clinician administered PTSD scale (CAPS): A Rasch Model analysis. *Journal of Applied Measurement, 4*, 59-69.

Bolelоucky, Z. & Horvath, M. (1974). The SCL-90 rating scale: First experience with the Czech version in healthy male scientific workers. *Activitas Nervosa Superior, 16*, 115-116.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum.

Brophy, C. J., Norvell, N. K., & Kiluk, D. J. (1988). An examination of the factor structure and convergent and discriminant validity of the SCL-90R in an outpatient clinic population. *Journal of Personality Assessment, 52*, 334-340.

Choquette, K. A. (1994). Assessing depression in alcoholics with the BDI, SCL-90-R, and DIS criteria. *Journal of Substance Abuse, 6*, 295-304.

Clark, A. & Friedman, M. J. (1983). Factor structure and discriminant validity of the SCL-90 in a veteran psychiatric population. *Journal of Personality Assessment, 47*, 396-404.

Cole, J.C., Rabin, A.S., Smith, T.L., & Kaufman, A.S. (2004). Development and validation of a Rasch-derived CES-D Short Form. *Psychological Assessment, 16*, 360-372.

Crits-Cristoph, P. (1992). The efficacy of brief dynamic psychotherapy: A meta-analysis. *American Journal of Psychiatry, 149*, 151-158.

Derogatis, L. R. (1975). *The SCL-90-R*. Baltimore, MD: Clinical Psychometric Research.

Derogatis, L. R. (1977). *SCL-90-R: Administration scoring and procedures manual*. Baltimore, MD: Clinical Psychometric Research.

Derogatis, L. R. (1994). *SCL-90-R: Administration, scoring and procedures manual*. Minneapolis, MN: National Computer Systems.

Derogatis, L. R. & Cleary, P. A. (1977). Confirmation of the dimensional structure of the SCL-90: A study in construct validation. *Journal of Clinical Psychology, 33*, 981-990.

Derogatis, L. R., Lipman, R. S., & Covi, L. (1973). SCL-90: An outpatient psychiatric rating scale – preliminary report. *Psychopharmacol Bulletin, 9*, 13-27.

Derogatis, L. R., Lipman, R. S., Rickels, K.,Uhlenhuth, E. H., & Covi, L. (1974). The Hopkins Symptom Checklist (HSCL): A self-report symptom inventory. *Behavioral Science, 19*, 1-15.

Derogatis, L. R., Rickels, K., & Rock, A. (1976). The SCL-90-R and the MMPI: A step in the validation of a new self-report scale. *British Journal of Psychiatry, 128*, 280-289.

Derogatis, L. R., & Savitz, K. L. (1999). The SCL-90-R, Brief Symptom Inventory, and matching clinical rating scales. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment, second edition (pp. 679-724).* Mahwah, NJ: Lawrence Erlbaum.

Dinning, W. D., & Evans, R. G. (1977). Discriminant and convergent validity of the SCL-90 in psychiatric inpatients. *Journal of Personality Assessment, 41*, 304-310.

Elliott, R., Clark, C., Wexler, M., Kemeny, V., Brinkerhoff, J., &Mack, C. (1990). The Impact of Experiential therapy of depression: Initial results. In G. Lietaer, J. Rombauts, & R. Van Balen (Eds.), *Client-centered and experiential psychotherapy towards the nineties* (549-577). Leuven, Belgium: Leuven University Press.

Elliott, R., Hitt, R., Klein, M., J., Partyka, R., Amer, M., Wright, et al. (June, 2002). *Quantitative Outcome of Process-Experiential Therapy in a Naturalistic Research Protocol*. Paper presented at meeting of the Society for Psychotherapy Research, Santa Barbara, CA.

Fischer, G.H. & Molenaar, I.W. (1995). Rasch models: Foundations, recent developments, and applications. New York, NY: Springer-Verlag.

Fisher, Jr., W. P. (1992). Reliability statistics. *Rasch Measurement Transactions* 6, 238.

Holcomb, W. R., Adams, N. A., & Ponder, H. M. (1983). Factor structure of the Symptom Checklist-90 with acute psychiatric inpatients. *Journal of Consulting & Clinical Psychology, 51*, 535-538.

Horowitz, L. M., Rosenberg, S. E., Baer, B. A., Ureno, G., & Villasenor, V. S. (1988). Inventory of interpersonal problems: Psychometric properties and clinical applications. *Journal of Consulting & Clinical Psychology, 56*, 885-892.

Kim, S. W., Dysken, W. W. (1990). Open fixed does trial of fluoxetine in the treatment of obsessive compulsive disorder. *Drug Development Research, 19*, 315-319.

Koeter, M. W. (1992). Validity of the GHQ and SCL-90-R anxiety and depression scales: A comparative study. *Journal of Affective Disorders, 24*, 271-279.

Kopta, S. M., Howard, K. I., Lowrey, J. L., & Beutler, L. E. (1994). Patterns of symptomatic recovery in psychotherapy. *Journal of Clinical and Consulting Psychology, 62*, 1009-1016.

Lambert, M. J., Burlingame, G. M., Umphress, V., Hansen, N.,Yanchar, S. C. Vermeersch, D., & Clouse, G. C. (1996). The reliability and validity of a new psychotherapy outcome questionnaire. *Clinical Psychology and Psychotherapy, 3*, 249-258.

Levine, S., Anderson, D., Bystritsky, A., & Barton, D. (1990). Eight HIV-seropositive patients with major depression responding to fluoxetine. *Journal of Acquired Immune Deficiency Syndrome, 3*, 1074-1077.

Linacre, J.M. (1996). True-score reliability or Rasch statistical validity? *Rasch Measurement Transactions, 9*, 455-456.

Linacre, J.M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*, 103 - 122.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*, 85-106

Linacre, J.M., & Wright, B.D. (2004). *WINSTEPS: Multiple-choice, rating scale, and partial credit Rasch analysis* [computer software]. Chicago, IL: MESA Press.

Lopez W. (1996). Communication validity and rating scales. *Rasch Measurement Transactions 10*, 482-3.

Low G.D. (1988). The semantics of questionnaire rating scales. *Evaluation and Research in Education 2*, 69-70.

Merbitz, C., Morris, J., & Grip, J.C. (1989). Ordinal scales and foundations of misinference. *Archives of Physical Medicine and Rehabilitation*, *70*, 308-312.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-759.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology,* 88(3), 355 - 383.

Mineka, S., Watson, D., & Clark, L. A. (1998). Comorbidity of anxiety and unipolar mood disorders. *Annual Review of Psychology, 49*, 377-412.

Mool, S. E. (1998). Alternative scoring and significant items of two popular self report inventories as related to global assessment of functioning measures. *Dissertation Abstracts International, 59(B)*, 0422.

Olsen, L. R., Mortensen, E.L., Bech, P. (2004). The SCL-90 and SCL-90R versions validated by item response models in a Danish community sample. *Acta Psychiatrica Scandinavica, 110*, 225-229.

Peveler, R. C. & Fairburn, C. G. (1990). Measurement of neurotic symptoms by self-report questionnaire: Validity of the SCL-90-R. *Psychological Medicine, 20*, 873-879.

Radloff LS. (2000). Center for Epidemiological Studies depressed mood scale [CES D]. In K. Corcoran & J. Fischer, *Measures for clinical practice: A sourcebook* (3rd Ed.) (Vol.2, Pp.155-154). New York, NY: Free Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danmarks. Paedagogiske Institut.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (expanded ed.). Chicago, IL: University of Chicago.

Shaw, F. (1991). Descriptive IRT versus Prescriptive Rasch. *Rasch Measurement Transactions, 5:1,* 131.

Smith, Jr., E. V. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement*, *2*, 281-311.

Smith, Jr., E.V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement, 3*, 205-231.

Smith, Jr., E.V., & Smith, R.M. (2004). *Introduction to Rasch measurement*. Maple Grove, MN: JAM Press.

Walsh, B. T., Hadigan, C. M., Devlin, M. J., Gladis, M., Roose, S. P., Fleiss, J., C., et al. (1997). Medication and psychotherapy in the treatment of bulimia nervosa. American *Journal of Psychiatry, 154*, 523-531.

Wright, B.D., & Masters, G.N. (1982). *Rating scale analysis.* Chicago, IL: MESA Press.

Wright, B. D., & Stone, M H. (2004). *Making Measures*. Chicago, IL: Phaneron Press.

Table 1. *Summary of the SCL-90-R 5-Point Rating Scale Category Functioning*

Category Label	Observed Count	Infit Mean Square	Outfit Mean Square	Step Threshold	Step Standard Error
0 (not at all)	13240	1.00	1.00	None	
1 (a little bit	6389	.93	.78	-.57	.01
2 (moderately)	3245	.93	1.01	-.03	.02
3 (quite a bit)	2357	1.09	1.32	.02	.02
4 (extremely)	1392	1.09	1.24	.57	.03

Note. *Observed count* includes all clients' responses for a category; *Infit Mean Square* measures deviation from measurement model for category and provides sensitivity to on-target (i.e., midrange) observations (1.0 is ideal; acceptable range: .6 – 1.4). *OutfitMean Square* measures deviation from measurement model for category and provides sensitivity to off-target, extreme responses (1.0 is ideal; acceptable range: .6 – 1.4); *Step Threshold* is the value on the logit transformed measure scale at which a response category becomes more probable than not (see Figure 1).

Table 2. *Summary of Changes in Person and Item Separation and Reliability as a Result of Collapsing Rating Scale Categories and Removing Misfitting Items*

Rating Scale	Separation (G)		Reliability		Infit Mean Square	Outfit Mean Square	# Misfitting Items
	Person	Item	Person	Item			
Original 5-point scale	4.90	7.76	.96	.98	1.02	1.05	11
4-point scale (combining 2 and 3)	5.07	8.35	.96	.99	1.00	1.02	4 (60, 84, 05, 20)
3-point scale (combining 1 and 2; and 3 and 4)	5.05	8.36	.96	.99	1.00	1.00	3 (60, 84, 05)
4-point scale (combining 2 and 3; and removing 4 misfitting items)	5.10	8.56	.96	.99	1.00	1.03	0
3-point scale (combining 1 and 2; 3 and 4; and removing 3 misfitting items)	5.07	8.52	.96	.99	1.00	1.00	0

Note. Separation (G) is the ratio of the modeled standard deviation to the standard error of measurement (including error due to misfit); for an explanation of Infit Mean Square and Outfit Mean Square, see note for Table 1. The alternative solutions were tried in the order presented here, guided by the goals of (a) maximizing separation and (b) retaining items but (c) reducing number of rating scale categories. The bottom row is the alternative used in the text.

Table 3. *Summary of the SCL-90-R New 3-Point Rating Scale Functioning*

Category Label	Observed Count	Infit Mean Square	Outfit Mean Square	Step Threshold	Step Standard Error
0 (not at all)	12794	1.00	1.00	None	
1 (a little bit + moderately)	9353	.95	.92	-.94	.02
2 (quite a bit + extremely)	3589	1.03	1.11	.94	.02

Note. For explanation column statistics, see footnote for Table 1.

Table 4. *Summary of Subscale Analysis*

Subscale	Person Separation	Person. Reliability	Item Separation	Item Reliabilty	Item Fit/ Misfit
Anxiety	1.72	.75	7.76	.98	All fit
Depression	2.29	.84	7.70	.98	20, 22, 5
Hostility	1.28	.62	9.90	.99	All fit
Interpersonal Sensitivity	2.01	.80	7.88	.98	All fit
Obsessive Compulsive	2.06	.81	7.38	.98	65
Paranoid Ideation	1.31	.63	3.40	.92	All fit
Phobic Anxiety	.72	.34	5.07	.96	75
Psychoticism	1.21	.60	8.22	.99	All fit
Somatization	1.71	.75	6.23	.97	All fit

Table 5. *List of Items on Which Two Sample Significantly Differed*

	Naturalistic		Depressed		**t**	ES
Item	Sample	Error	Sample	Error		
1. Pains in head	-0.47	0.13	-1.35	0.14	4.61	0.53
20. Weeping readily	0.48	0.15	-0.38	0.14	4.19	0.49
67. Destroy objects	1.16	0.17	0.31	0.16	3.64	0.42
84. Lewd ideas	1.04	0.17	0.26	0.15	3.44	0.40
34. Overly sensitive	-0.73	0.13	-1.35	0.14	3.25	0.38
31. Excessive concern	-1.38	0.13	-2.01	0.15	3.17	0.37
12. Pangs upper body	1.15	0.17	0.41	0.16	3.17	0.37
75. Anxious by self	1.19	0.17	0.47	0.16	3.08	0.36
44. Insomnia	-0.78	0.13	-0.28	0.14	-2.62	-0.30
21. Other gender	-0.39	0.13	0.13	0.15	-2.62	-0.31
5. Lack erotic	-0.69	0.13	-0.13	0.15	-2.82	-0.33
13. Fear exposed	0.86	0.16	1.68	0.22	-3.01	-0.36
18. Others unreliable	-0.78	0.13	-0.18	0.15	-3.02	-0.35
52. Loss feeling	-0.03	0.14	0.82	0.17	-3.86	-0.45
22. Ensnared	-0.52	0.13	0.29	0.16	-3.93	-0.46

Note. Results of Rasch DIF analysis comparing logit scores across samples; positive t-values indicate higher levels in the naturalistic sample. All reported *t* values are significant at $p < .01$.

Figure Captions

Figure 1. Analysis of SCL-90-R 5-point rating scale categories: Probability of response categories as a function of adjusted client distress. Adjusted Clinical Distress is client distress minus item difficulty (both expressed as logit scores); Probability of Category is the likelihood of endorsing a given rating scale category at that level of Adjusted Clinical Distress. Intersection of adjacent rating scale categories can be seen at estimated threshold value of the higher of the two categories. For example, the threshold value for category 1 is -.57 (reported in Table 1 and visually represented in this figure); the probability of choosing category 1 at this level is slightly less than .4, as shown as the height of the intersection on the y axis. Figure generated using WINSTEPS 3.57 (Linacre & Wright, 2004).

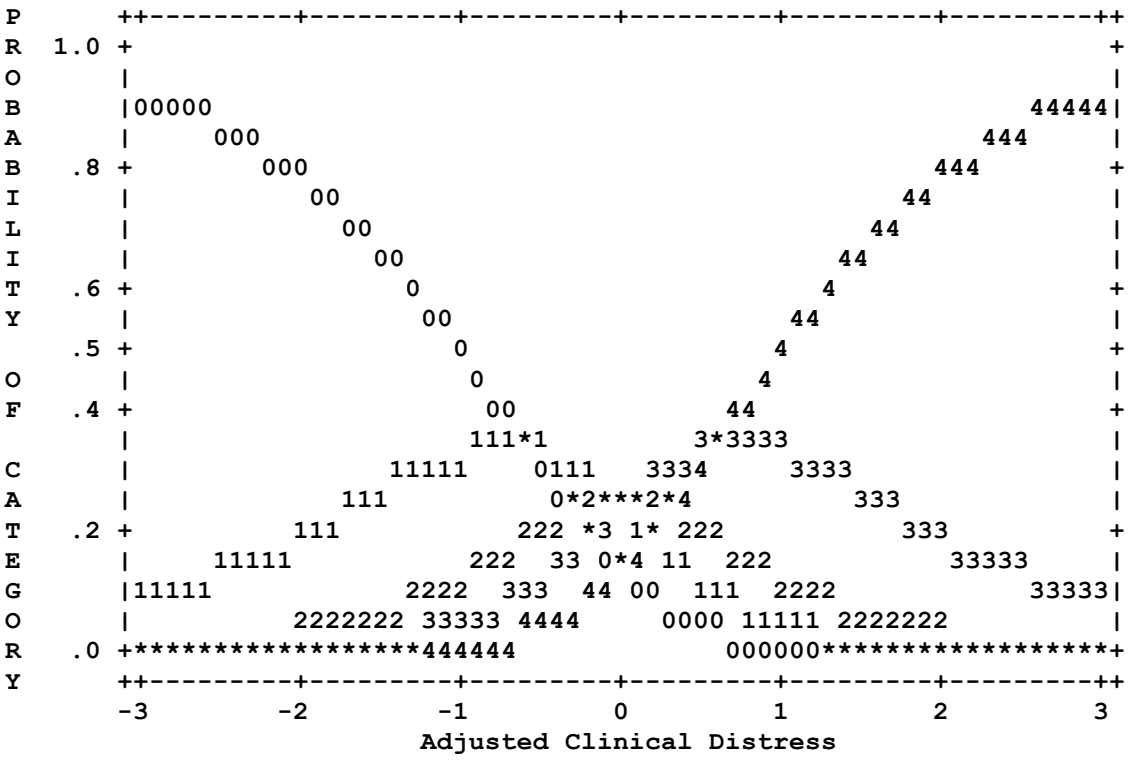


Figure 2. Analysis of modified SCL-90-R 3-point rating scale categories: Probability of response categories as a function of adjusted client distress. Intersection of adjacent rating scale categories is shown at estimated threshold value of the higher of the two categories. For example, the threshold value for category 1 is -.94 (obtained from Table 3 and visually shown in this figure) The probability of choosing category 1 at the threshold is slightly less than .5, as shown as the height of the intersection on the y axis. Figure generated using WINSTEPS 3.57 (Linacre & Wright, 2004).

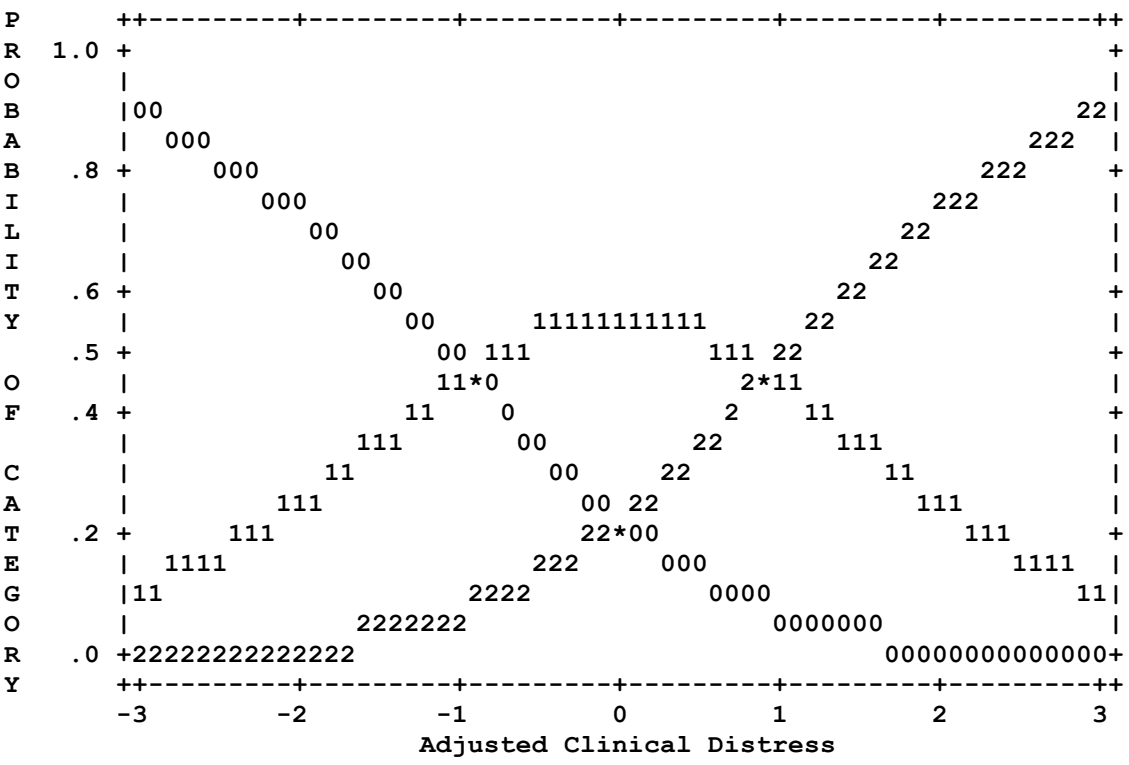
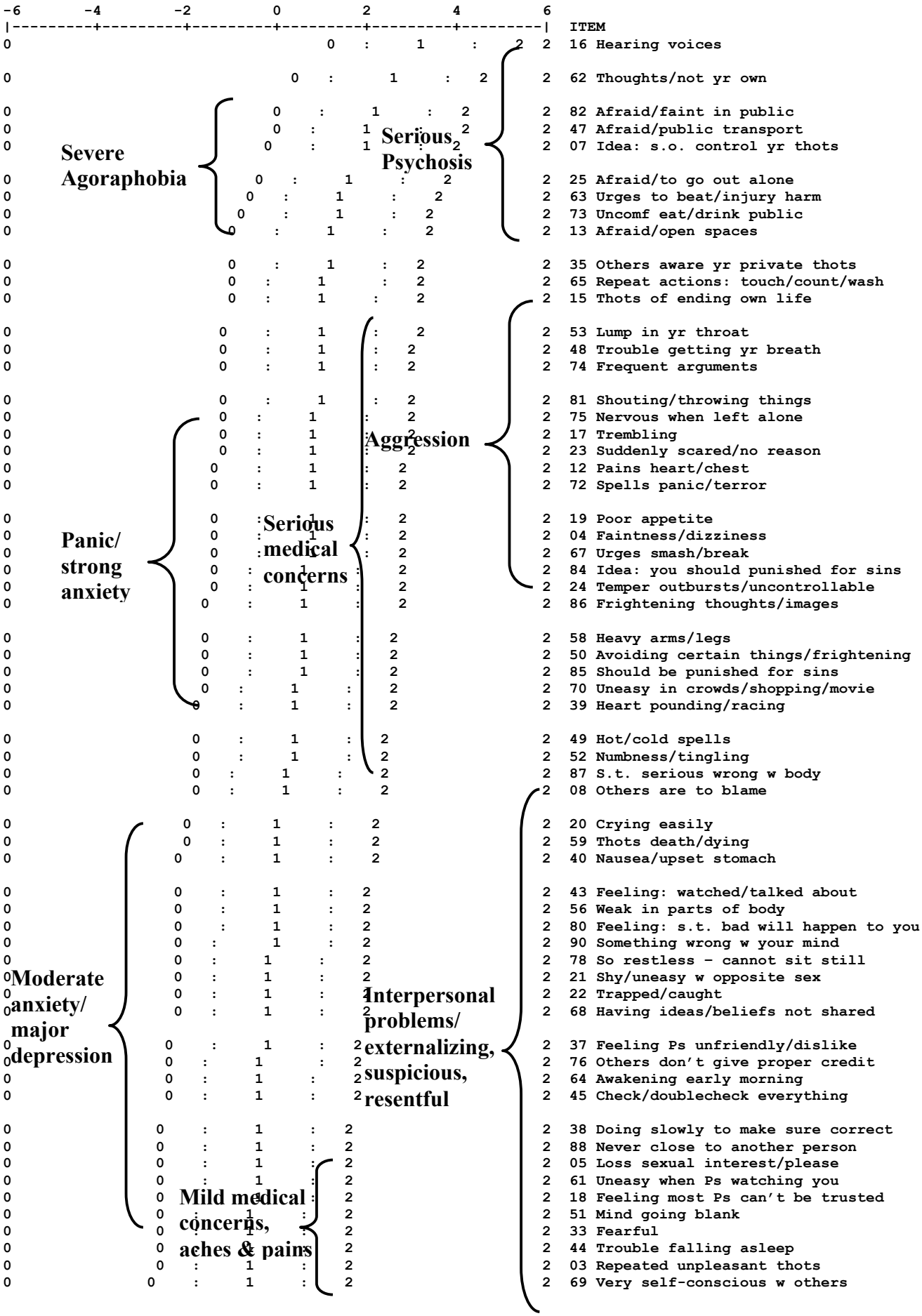
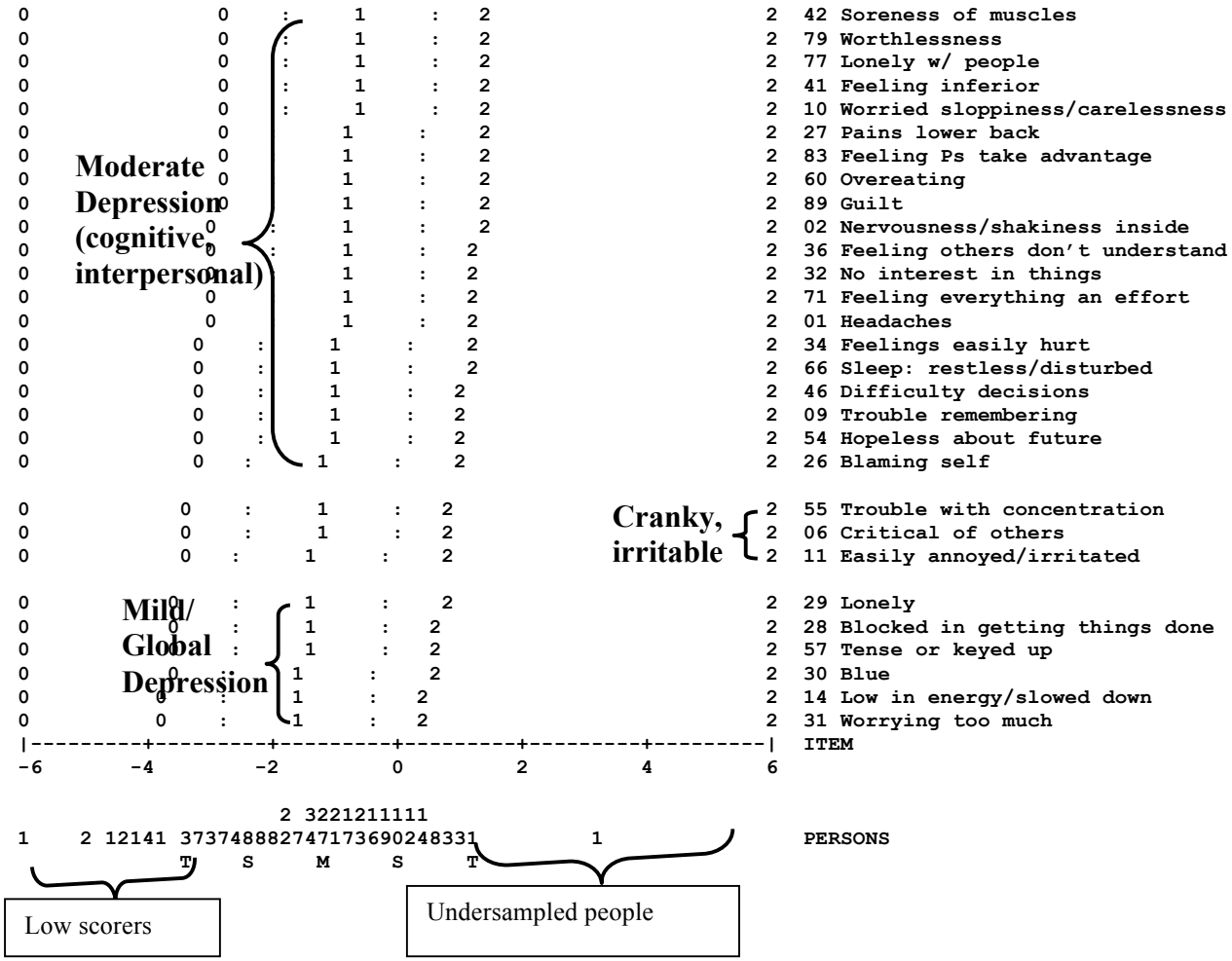


Figure 3. Person-item map. Items are abbreviated and paraphrased from the SCL -90-R® (Symptom Checklist–90–Revised), Copyright © 1975 Leonard R. Derogatis. Adapted with permission from Pearson Assessments, Minneapolis MN. For an accurate interpretation, consult the instrument (e.g., Derogatis, 1994). The items are listed from those showing most clinical distress (Figure 3; top of the map) to least (Figure 3, cont.; bottom of the map). The body of the figure shows the estimated category responses (with the collapsed 3 point scale) for each item, based on a person's position on the measure (x axis). At the bottom is the person distribution, frequencies given by vertically-stacked numbers; M = mean; S = 1 *sd* from mean; T = 2 *sd* from mean. Figure generated using WINSTEPS 3.57 (Linacre & Wright, 2004), with added annotations.


-6 -4 -2 0 2 4 6
ITEM
16 Hearing voices
62 Thoughts/not yr own
82 Afraid/faint in public
47 Afraid/public transport
07 Idea: s.o. control yr thots
25 Afraid/to go out alone
63 Urges to beat/injury harm
73 Uncomf eat/drink public
13 Afraid/open spaces
35 Others aware yr private thots
65 Repeat actions: touch/count/wash
15 Thots of ending own life
53 Lump in yr throat
48 Trouble getting yr breath
74 Frequent arguments
81 Shouting/throwing things
75 Nervous when left alone
17 Trembling
23 Suddenly scared/no reason
12 Pains heart/chest
72 Spells panic/terror
19 Poor appetite
04 Faintness/dizziness
67 Urges smash/break
84 Idea: you should punished for sins
24 Temper outbursts/uncontrollable
86 Frightening thoughts/images
58 Heavy arms/legs
50 Avoiding certain things/frightening
85 Should be punished for sins
70 Uneasy in crowds/shopping/movie
39 Heart pounding/racing
49 Hot/cold spells
52 Numbness/tingling
87 S.t. serious wrong w body
08 Others are to blame
20 Crying easily
59 Thots death/dying
40 Nausea/upset stomach
43 Feeling: watched/talked about
56 Weak in parts of body
80 Feeling: s.t. bad will happen to you
90 Something wrong w your mind
78 So restless - cannot sit still
21 Shy/uneasy w opposite sex
22 Trapped/caught
68 Having ideas/beliefs not shared
37 Feeling Ps unfriendly/dislike
76 Others don't give proper credit
64 Awakening early morning
45 Check/doublecheck everything
38 Doing slowly to make sure correct
88 Never close to another person
05 Loss sexual interest/please
61 Uneasy when Ps watching you
18 Feeling most Ps can't be trusted
51 Mind going blank
33 Fearful
44 Trouble falling asleep
03 Repeated unpleasant thots
69 Very self-conscious w others
Severe Agoraphobia
Serious Psychosis
Aggression
Serious medical concerns
Panic/ strong anxiety
Moderate anxiety/ major depression
Interpersonal problems/ externalizing, suspicious, resentful
Mild medical concerns, aches & pains


Moderate Depression (cognitive, interpersonal)
Cranky, irritable
Mild/ Global Depression
2 42 Soreness of muscles
2 79 Worthlessness
2 77 Lonely w/ people
2 41 Feeling inferior
2 10 Worried sloppiness/carelessness
2 27 Pains lower back
2 83 Feeling Ps take advantage
2 60 Overeating
2 89 Guilt
2 02 Nervousness/shakiness inside
2 36 Feeling others don't understand
2 32 No interest in things
2 71 Feeling everything an effort
2 01 Headaches
2 34 Feelings easily hurt
2 66 Sleep: restless/disturbed
2 46 Difficulty decisions
2 09 Trouble remembering
2 54 Hopeless about future
2 26 Blaming self
2 55 Trouble with concentration
2 06 Critical of others
2 11 Easily annoyed/irritated
2 29 Lonely
2 28 Blocked in getting things done
2 57 Tense or keyed up
2 30 Blue
2 14 Low in energy/slowed down
2 31 Worrying too much
ITEM
-6 -4 -2 0 2 4 6
2 3221211111
1 2 12141 37374888274717369024833 1
PERSONS
T S M S T
Low scorers
Undersampled people