Chapter 2

Test Development Procedures:
Psychometric Study and Analytical Plan

Introduction

The Peabody Treatment Progress Battery (PTPB) is designed to provide clinically relevant measures of key mental health outcomes and clinical processes. The measures, especially with their repeated use, offer clinicians systematic feedback on their clients, both individually and in relation to other clients served. Such feedback provides rich clinical material for treatment planning, particularly for clients with outcomes that are not improving as expected.

A useful test battery has the following qualities:

- Reliable: Every question in a battery must contribute accurate information.
- Valid: Scores must have evidence-based interpretations.
- Short: Measures must be feasible to administer in the time available. A brief battery enables clinicians and clients to spend their time more effectively.
- Theory-based: A battery must have an understandable theoretical core so that clinicians, caregivers, and youths can understand the results.
- Integrated: A battery must cover main issues in a cohesive, integrated way, something a collection of unrelated instruments cannot do.

We believe that the PTPB meets these criteria. After two independent and comprehensive psychometric studies, we have strong evidence of the reliability of the instruments. We also have some evidence for the battery’s construct validity including estimates of factorial and discriminant validity for all instruments, as well as estimates of concurrent (i.e., convergent) validity for the Symptoms and Functioning Severity Scale (SFSS) from the first edition of the PTPB in the form of correlations with widely used instruments, namely the Child Behavior Checklist (CBCL; Achenbach, 1991) and Youth Self Report (YSR; Achenbach); the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1999); the Youth Outcome Questionnaire (Y-OQ®; Wells, Burlingame, & Lambert, 1999) and the Child and Adolescent Functioning Assessment Scale (CAFAS; Hodges & Wong, 1996). We think that the instruments in the PTPB have good reliability and evidence of validity, but it will take more research to further evaluate their construct validity (Cronbach, 1955) and predictive validity for monitoring the process and outcome of counseling.
An objective multi-method approach was used to assure that the PTPB met the criteria described above, including expert review, cognitive testing, psychometric study, and a rigorous analysis plan.

This second edition of the PTPB builds upon the 2007 version with a second round of intensive and comprehensive psychometric testing and measure revision. The current edition also includes a new measure chapter dedicated to the Session Report Form (see chapter 14). We believe the current edition of the PTPB represents an even stronger battery of clinically relevant measures revised to be as brief as possible while maintaining strong psychometric qualities and still being able to provide important information on key mental health outcomes and clinical processes.

**Development of the First Edition of the PTPB**

Below is a brief summary in the development of the first edition of the PTPB, published in 2007. The remaining chapter describes the extensive development process that has taken place since then and has resulted in this second edition.

**Expert Reviews**

The measures were distributed to a large number of clinicians, researchers, and other experts in the field who provided feedback on the usefulness and validity (face, content) of measures and of items within each measure. This resulted in some revisions (e.g., items reworded, added, dropped).

**Cognitive Testing**

The instruments were then administered to convenient samples of youths and their adult caregivers. Trained staff conducted separate cognitive interviews with youths and caregivers to help assure the validity of measures and ensure age appropriate readability.

**Psychometric Study**

We conducted a study in 2005 to obtain information about youths served, to test the psychometric properties of measures in the PTPB, and to get feedback from clients, adult caregivers, and clinicians on their perceived utility of the measures. The sample for this study was made available to us by a national social service agency (our collaborating partner in CFS™) that delivers home- and community-based services for youths and their families. In 2006, the agency served over 43,000 clients with 600 contracts in 32 states and the District of Columbia. Over 600 youths in 28 offices (called regions by the service organization), mostly located in the Eastern or Midwestern U.S., participated in the study.

**Rigorous Analysis Plan**

Analyses of the first edition of the PTPB assessed reliability, validity, and brevity to produce an integrated set of measures that can be administered efficiently and at low cost. Items were selected using objective public criteria from classical test theory, exploratory and confirmatory factor analysis, and Rasch modeling (Bond & Fox, 2001; Linacre & Wright, 2006), a single-parameter member of the IRT (Item Response Theory) family.
All measures were reduced to the minimum length consistent with traditional reliability (Cronbach & Shavelson, 2004) and person-separation reliability. The temporal stability and concurrent validity of the essential mental health measure (SFSS) were evaluated.

**Current Psychometric Study**

A description of the general analytic procedures is below. The history of development, theory base, and statistical results for each instrument are presented in chapters 4 to 13 and 15. The examination of the measurement quality of all instruments in the PTPB is continuing, with future updates planned for this manual. This second edition of the PTPB includes measure revisions and updated psychometrics from a second round of rigorous psychometric testing conducted in 2009 as well as the introduction of the Session Report Form (SRF; Chapter 14). The psychometric sample was drawn from a larger study evaluating the effects of CFS on clinician behavior and client outcomes sponsored by the National Institute for Mental Health.

**Eligibility Criteria**

Eligible youths were 11-18 years old; they were receiving mental health services; and their clinicians thought they could understand questions in the PTPB. One “primary” adult caregiver was also asked to participate if any was present at the time forms were completed. The sample included participants (youths, caregivers and clinicians) from 38 different sites of a national behavioral health organization. All clinicians were eligible to participate along with all of their adolescent clients already in services or who presented for services during the study period. If a youth had more than one clinician, the one considered the primary clinician and who saw the youth during the data collection period completed the clinician forms.

**Confidentiality**

All data were obtained as part of our collaborating partner’s continuous quality improvement (CQI) initiative; Contextualized Feedback Intervention and Training (CFIT) (see Chapter 15 for more information on CFIT). This system has been replaced by a new and much improved system called Contextualized Feedback Systems® or CFST™ (CFsystemsOnline.com). Vanderbilt had no contact with participants either for recruitment or data collection. Names or other information that could readily identify respondents were not sought or obtained. All data received and maintained by Vanderbilt included only a unique non-sensitive ID number for each participant.

**Administration Protocols**

Sometimes clinicians need to assist youths or caregivers whose reading is impaired. For the study, clinicians were provided their own copy of all youth and caregiver questionnaires that they could read to youths and caregivers, as needed. Clinicians were to read the instructions first, then the question, then the possible answer choices, giving respondents enough time to mark their answers on their own forms. Clinicians could repeat answer choices if needed. Youths/caregivers were asked to answer items the best way they could, as they understood them. Clinicians were instructed to help, if requested,
with questions but not to help with answers. We would recommend protocols such as these to PTPB users.

In order to help assure widespread use of the battery, all instruments were written for respondent understanding at the 4th grade reading literacy level. The literacy level was tested with Microsoft Office Word and then with the online SMOG (simplified measure of gobbledygook) Calculator. Standard forward-backward translation methods were used to create Spanish versions. A bilingual, native Hispanic translated the original English version into Spanish. A second bilingual, non-native speaker back-translated the Spanish version into an English version, which was compared to the original English version to assure accuracy of translation.

**General Procedures**

Each youth entered into the feedback system (CFST™; as described in chapter 15 as part of the larger study) received a questionnaire schedule containing a combination of PTPB measures for each week, as well as for baseline and discharge assessments that were one-time, non-repeating packets. The questionnaires scheduled for each week—according to the pre-programmed scheduled—contained a combination of youth, caregiver, and clinician questionnaires that could be printed from the system. The questionnaires were taken to each youth’s treatment session. The youth, a primary adult caregiver, if present, and clinician completed the measures at the end of a session to reduce any undue influence completing the forms might have on the therapeutic process. Clinicians were allowed to read questions to youths and adult caregivers, but were instructed not to help with answers. All youth and adult caregiver measures were available in English and Spanish. After all questionnaires had been completed, all respondents placed their questionnaires into a large envelope that the clinician sealed and then signed across the seal for confidentiality purposes. Clinicians returned their sealed and signed envelope to their office for data entry. The data entry staff were typically administrative assistants or office managers. Questionnaires were reviewed for completion and accurate answers (e.g., did not draw picture with answer responses), and then data entered into the feedback system. Once data entered, questionnaires were scored according to their individual psychometrics and an online feedback report became available. The database was translated into SAS system files. Univariate statistics (e.g., frequencies; means) for each variable were generated and examined for accuracy.

Within the year-long data collection of the larger study evaluating the effects of CFST™ on clinician behavior and client outcomes sponsored by the National Institute for Mental Health (Bickman, Kelley, Breda, Vides de Andrade, & Riemer, 2011), youth entered and left treatment (and thus the feedback system) at different times. Therefore, the larger dataset is dynamic as the sample changes throughout time. The sample utilized for current psychometric purposes represents a static snap-shot within this dynamic dataset and includes the first time each respondent completed each respective measure. Therefore, the sample sizes vary from measure to measure.
As Table 2.1 shows, a total of 749 youths, 695 caregivers, and 301 clinicians contribute data for the psychometric sample. See individual measure chapters (chapter 4-14) for measure-specific sample sizes included in analyses.

**Table 2.1  Sample Size by Respondent**

<table>
<thead>
<tr>
<th>Respondent</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Youths</td>
<td>749</td>
</tr>
<tr>
<td>Adult Caregivers</td>
<td>695</td>
</tr>
<tr>
<td>Clinicians</td>
<td>301</td>
</tr>
</tbody>
</table>

**Sample Characteristics**

Of the 749 youths in the sample, 46% were female. Approximately a third (30%) were African American; 49% white; 9% of other ethnic background and 12% multi-ethnic. Seventeen percent was Hispanic. On average, youths were 15 years old, with 15% between 11-12 years of age; 47% between 13-15 years of age, and 38% between 16-18 years of age. Based on clinician reports, the majority (89%) of youths had been diagnosed with a mental health disorder at some time before or with current treatment. Almost half (45%) of the youths had used alcohol or drugs.

On average, caregivers were 45 years old. Approximately 86% were female. About 33% were African American; 57% white; and 10% multi-ethnic or of other ethnic backgrounds. Almost twelve percent was Hispanic. About 22% of caregivers had not received a high school diploma; nearly two-thirds (60%) had completed high school; and 17% had earned a bachelor’s or higher degree. Almost half (44%) had annual household incomes less than $20,000; 31% had incomes of $35,000 or more. About a third of caregivers were divorced or separated, about half were currently married or living as married, 15% were never married, and others were widowed.

Nearly all (96%) caregivers considered themselves to be the youth’s primary caregiver. About half (55%) were the youth’s birth parent, 19% were a foster parent, another 11% grandparents and others included other family members. The majority (89%) reported they knew the youth at least pretty well, with others (virtually all foster parents) reporting they did not know the youth too well.

**Analytical Plan**

The psychometric analyses emphasized evaluating every item of every instrument for its reliability and validity.

**Determining Reliability**

Reliability of each instrument in the battery was determined using multiple methods, such as Cronbach’s alpha, item-total correlations, and Rasch person-separation reliability. Additional methods used to minimize test length are described below. At this point, not
all psychometric measures of reliability are available for all instruments (see chapters 4-13 & 15 for a discussion of tests for each measure).

**Reliability Coefficients**

Cronbach’s alpha internal consistency reliability estimates were calculated for total scores and any subscale scores. Cronbach’s alpha is higher when internal consistency is high and smaller when it is low. Alpha also increases with test length (Brown, 1910; Spearman, 1910). A Cronbach’s alpha of 0.80 or higher is generally considered satisfactory (Nunnally & Bernstein, 1994). Satisfactory Cronbach’s alpha indicates that a measure is of sufficient length and that the items appear to be measuring similar content.

**Comprehensive Item Psychometrics**

We examined each item using currently available models for psychometrics, namely classical test theory (CTT), confirmatory factor analysis (CFA), and Rasch modeling. We view all three methods as useful tools, each with strengths and limitations for creating brief instruments for frequent use. Each of the models produces information to identify stronger and weaker items in a given test. By putting this information into a single table, we can evaluate a test and its items at a glance. Information about the statistical merits of each item is necessary to determine whether a test should be revised. Note that throughout this manual, we use the terms Rasch modeling and Item Response Theory (IRT) interchangeably to refer to the logistic model-based approach to test development as compared to the CTT.

Table 2.2 shows the statistical information for a test or questionnaire with Likert-type scale items, and the criteria used to check each item.

**Table 2.2 Statistical Properties of Effective Test Items**

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Sought Values</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Between 2 and 4 (5 pt scale)</td>
<td>Avoid floors and ceilings in the target sample</td>
</tr>
<tr>
<td>Rasch Measure Score</td>
<td>Cover the range</td>
<td>Need items across the range of youth</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>Not extremely high</td>
<td>Avoid items where everyone gives same response</td>
</tr>
<tr>
<td>Item-Total Correlation.</td>
<td>Higher better</td>
<td>Keep items that measure a single thing</td>
</tr>
<tr>
<td>Infit &amp; Outfit</td>
<td>Between 0.5 and 1.5</td>
<td>Keep items that fit 1PL (logistic) model</td>
</tr>
<tr>
<td>Discrimination</td>
<td>Avoid low discrimination</td>
<td>Avoid items that can’t discriminate</td>
</tr>
</tbody>
</table>

Note: IRT with WINSTEPS® 3.63.0 (Linacre, 2007)

Below is a brief explanation of the rationale for each criterion:

- **Mean**: If an item’s mean is at the top or bottom of its range (e.g., a mean of 4.8 or 1.2 on a 1-5 scale) this means that nearly everyone is giving the same rating, making it impossible for that item’s meager variance to correlate with anything.
• **Kurtosis**: A good way to spot items with poor measurement ability is to focus on those that are excessively leptokurtic, meaning that a great number of people all have the same score. An example is the item “I have attempted suicide” in a nonclinical sample. Since nearly all would say “No,” the kurtosis is very high. Such items with extreme means and kurtosis have floor or ceiling artifacts that make little psychometric contribution to a test with normal subjects.

• **Rasch Measure Score**: An item’s difficulty or rarity as expressed in the measure score in the Rasch logistic model shows where the item is efficient and informative about a given test taker. For example, an “easy” item to endorse such as “I have worried more than once” might tell us nothing about differences between serious cases of psychopathology. The most efficient strategy for accurate measurement is to have a range of items from very easy to very difficult or unusual (e.g., “I have committed homicide”) so that the entire range of clients is measured reliably. Rasch measure scores were scaled to have a mean of 0 and a standard deviation of 1.

• **Item-Total Correlation**: In classical test theory (Lord, Novick, & Numerous, 1968), test designers realized that if a group of items is a reliable measure of a single construct, such as psychopathology or intelligence, the items must be correlated. By observing the correlation between each item and the sum of all the others, we can identify the odd unrelated items by their low correlation. It is senseless to add unrelated items into an index score. By dropping items with low item-total correlations, we create an index with high internal consistency.

• **Rasch Model Infit and Outfit**: Since the Rasch model defines good measurement, items that fit the model are good items, and scores on the good items show a consistent s-shaped logistic to give relationship to the person’s strength of the measured trait. The Infit mean square measures model fit for the middle cases in the distribution; Outfit, for the extreme cases at the tails of the distribution. According to Linacre (2006), items with an Infit and Outfit between 0.5 and 1.5 (Linacre, 2006) contribute to the reliability of measurement and items outside that range do not. When possible, we follow the stricter criteria of 0.7 – 1.3 (Bond & Fox, 2001).

• **Rasch Model Discrimination**: While the Rasch model is a 1-parameter logistic model, WINSTEPS® 3.63.0 (Linacre, 2007) estimates each item’s discrimination after the 1 parameter logistic model is estimated. Items with low discrimination are less effective at detecting which people are high or low on the trait measured.

One goal in this second edition was to ensure our measures were as brief as possible while maintaining strong psychometric properties and providing maximum information. Thus, a large focus was determining whether items could be dropped from some instruments based on all of these indices of item quality. Please understand that while these criteria provide guidance, there are no agreed upon cut scores for most of them. Fortunately, in our testing, the eliminated items had warning flags on multiple items.
**Standard Errors of Measurement**

The standard error of measurement (SEM) shows how much uncertainty there is around each youth’s score on a given occasion. This statistic is smaller (i.e., more precise) when reliability is high, and larger when it is low (i.e., the test’s standard deviation is high). Assuming, for example, that the total score on the adult caregiver SFSS (full version) was 60, a standard error of 2.50 would mean that the true severity level for the youth is between 55.00 and 65.00 with approximately 95% certainty. Thus, the smaller the standard error, the more precise the measurement.

**Minimum Detectable Change**

The index of minimum detectable change (MDC; Schmitt & Di Fabio, 2004) provides guidance about when a client’s change in scores is so large that we can be fairly certain that the client either improved or got worse significantly. For example, if a youth’s self report score on the SFSS decreased by more than 4.63 points (which corresponds to the minimum detectable change threshold reported in each chapter), we would be 75% confident that the change in scores represents a significant improvement in the client’s severity level. This does not mean that a change that is less than the indicated reliable change threshold reported for each scale is not meaningful. However, because respondents to questionnaires do not always use the available answer options consistently (e.g., one time a youth uses “sometimes” and another time “often” to describe the frequency with which he got into trouble, even though nothing has really changed), we cannot be very certain whether the score change on the scale is related to a real change in severity level, if the change is less than the minimum detectable change threshold. If the change in scores is less than the MDC threshold, the clinician may want to determine in some other way if a change in scores represents real change. Reviewing which items contributed the most to the change in scores may provide some insight. It is also helpful to look at a trend with more than just two data points. A trend is more reliable than a simple change score, and provides information on whether a score is stable, improving, or declining over time.

**Test-Retest Reliability**

Test-retest reliability is a good indicator of an instrument’s temporal stability. This form of reliability is only available for the SFSS from the first edition of the PTPB. These results are reprinted below. The procedures for test-retest reliability for the SFSS are as follows.

Four different regional offices were targeted for a reliability test-retest of the SFSS. These regions received only the full SFSS and a few items on demographic background in two concurrent sessions. Additionally, for two of the regions the order of items in the SFSS was reversed between the two regions in order to test for order effects. Roughly the first half of items was at the beginning of the measure for one region (both administrations) while at the end of the measure for the other region (both administrations). There was no significant order effect in this sample.
between administrations were not included in this assessment of test-retest reliability. The average interval between the two administrations for the final sample was 7 days ($SD = 3.8$). Table 2.3 shows the number of clients in each region about whom youths, adult caregivers, and clinicians reported during the first (test) and second (retest) administrations of the SFSS.

Table 2.3  SFSS Reliability Test: Participation by Region

<table>
<thead>
<tr>
<th>SFSS</th>
<th># of Regions</th>
<th># of Youth Reports</th>
<th># of Adult Caregiver Reports</th>
<th># of Clinician Reports</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1&lt;sup&gt;st&lt;/sup&gt; Adm.</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt; Adm.</td>
<td>1&lt;sup&gt;st&lt;/sup&gt; Adm.</td>
</tr>
<tr>
<td>Reliability Test</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total for Reliability Test</td>
<td>4</td>
<td>105</td>
<td>84</td>
<td>85</td>
</tr>
</tbody>
</table>

Determining Validity

Factorial Validity

We evaluated the factorial validity of all PTPB measures. Factorial validity examines whether the test’s confirmatory measurement model fits the theory of what the test purports to measure. For example, theory suggests that the SFSS unlike Beck Depression, should not fit a unidimensional model because the SFSS is a broad sampling of different internalizing and externalizing problems. Based on the factorial structure validated in the first edition of the PTPB, the current version represents further validation of previously determined factor structure for model fit.

Scree Plots

A scree plot of eigenvalues is used in exploratory factor analysis for a visual inspection of whether the data are best represented by more than one factor. The plot shows the eigenvalue of each principle component. Scree plots (Cattell, 1966) have long been used to see if a test fits a one-factor model. If the second eigenvalue is less than two or three, then a robust second factor might not be possible. Some instruments, such as the Caregiver Strain Questionnaire- Short Form 7 (CGSQ-SF7), are multifactorial, meaning they have subtests. The multiple factors are apparent by inspection of scree plots, which are confirmed by confirmatory factor analysis.

Scree plots are tools of exploratory factor analyses. Whereas exploratory factor analysis (EFA) gives useful descriptions, confirmatory factor analysis (CFA) is provides more definite conclusions. CFA gives a direct answer whether the observed data fit an explicit theoretical model (Floyd & Widaman, 1995). The PTPB uses both EFA and CFA for what each method can contribute.

---

<sup>4</sup> Results reported are reprinted from the first edition of the PTPB (Bickman et al., 2007)
Confirmatory Factor Analysis

Confirmatory factor analysis (CFA) estimates how well a measurement model fits the data. For example, when several sub-scales are suggested by theory, they can be tested to see how well the theory fits the data. CFA lets us determine the “factorial validity” of each scale. While this form of validity is less important than criterion validity, it is necessary for the interpretation of scores. To estimate model fit, we used three popular fit statistics named in Table 2.4: Bentler’s Comparative Fit Index (CFI; Bentler, 1990), Joreskog’s Goodness of Fit Index (GFI; Joreskog, 1988), and the Standardized Root Mean Square Residual (SRMR; Steiger, 2000). A one-factor model required these rather exacting cutoffs (Yu, 2002). According to Browne and Cudeck (1993), values greater than 0.90 indicate good fit between a model and the data for the CFI and GFI. For the SRMR, a value of 0.05 indicates close fit, 0.08 fair fit, and 0.10 marginal fit (Hu & Bentler, 1999).

Table 2.4  Fit Criteria

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bentler’s CFI</td>
<td>0.90 or more</td>
</tr>
<tr>
<td>Joreskog’s GFI</td>
<td>0.90 or more</td>
</tr>
<tr>
<td>SRMR</td>
<td>Well fitting: 0.05 or less</td>
</tr>
<tr>
<td></td>
<td>Satisfactory: 0.08 to 0.051</td>
</tr>
</tbody>
</table>

Convergent and Discriminant Validity

Because these forms of validity are particularly important for a battery of instruments, an entire chapter is devoted to a discussion of the procedures and results (see Chapter 3).

General Analytical Issues

Age, Gender, Race and Ethnicity

For each instrument, multi-factorial ANOVAs, followed by multiple comparisons, were conducted to determine if any differences were apparent by youth age, gender, and racial or ethnic group. Significance levels were adjusted to account for the multiple tests using the Bonferroni adjustment. Several of the youth measures (TOES, BMSLSS-PTPB, SFSS) displayed a small gender effect with males scoring somewhat lower than females. These differences were so small in magnitude in the overall context of non-significant comparisons that it was decided not to present different norms for these groups at this time. It is also important to note that these analyses were based on a sample of youth who received services for their mental health problems. It may well be that there are more pronounced differences among these groups (gender, race, ethnicity, and age) in nonclinical population samples. Future analyses with additional samples are needed to determine whether different norms for these groups should be established.
Missing Data
Analyses for each measure were based on cases with 85% or more completed data on the measure. Among these cases, subject mean substitution was used for any missing data. Cases with less than 85% completed data were excluded entirely. Please note that the self-scoring forms provided with this battery can be completed directly for a subject when all data are present. When data are missing, users will need to compute the subject’s mean across all items in that measure first and substitute it for any missing item. Determining the cut point (e.g., 85%) for when too much data are missing to compute summary scores in this way is at the discretion of the user.

Instrument Scoring and Interpretation
In each chapter, specific scoring procedures are presented for the total score and any subscale scores (if applicable) with examples of the self-scoring forms and information on the clinical interpretation of scores. All PTPB measures and self-scoring forms can also be found in Appendix B: Measures and Self-Scoring Forms.

In addition to the reliable change index described previously, total and subscale scores for an individual respondent can be compared for that same respondent over time as a trend, and in relation to the psychometric sample. Comparisons to the psychometric sample are given in the form of low, medium, and high scores. The low, medium, high scores help to judge whether a score should be considered relatively low, medium, or high, compared to the current psychometric sample. For example, if a youth’s SFSS Total Score falls in the low range, it means that this youth is rated to have relatively low signs of symptoms and functioning relative to the youths in the psychometric client sample (not a normal school population). Percentile rank tables that represent the distribution of scores in the psychometric sample are presented as well. Additional samples for comparison will be available in the future.

Backward Compatibility to First Edition
All individual items included in the measures found in this second edition of the PTPB are also found on first edition measures. Although some items were dropped for the current edition, scale psychometrics are comparable across editions. Therefore, with the exception of the transformed scores of the SFSS measures, total scores from the current version of the PTPB measures are backward compatible and thus comparable with those from the first edition. This eases the transition of first edition users to the current updated edition by allowing individuals’ scale scores are able to be compared across editions.
References


