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 think the PTPB meets these criteria. At this time, we have strong evidence of the reliability of the instruments. We have some evidence for the construct validity including estimates of factorial and discriminant validity for all instruments, and estimates of concurrent (i.e., convergent) validity for the Symptoms and Functioning Severity Scale (SFSS) 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.

A 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.
**Expert Reviews**

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 from May to September, 2005 to obtain descriptive 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 CFIT) 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 as described below.

**Rigorous Analysis Plan**

Analyses of the PTPB assessed reliability, validity, and brevity to produce an integrated set of measures that can be administered efficiently and at low cost. Each measure was tested with classical test theory, exploratory and confirmatory factor analysis, and Rasch modeling. All measures were reduced to the minimum length consistent with reliability. Too, the temporal stability and concurrent validity of the essential mental health measure (SFSS) were tested. 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. The examination of the measurement quality of all instruments in the PTPB is continuing, with future updates planned for this manual.

**Psychometric Study**

**Eligibility Criteria**

Youths 11-18 years old inclusively whose clinicians thought they could understand questions such as those in the PTPB were eligible to participate in the study. One “primary” adult caregiver was also asked to participate if any was present at the time forms were completed. Clients had to have had at least one therapeutic session prior to study participation to help assure adequate familiarity among the youth, adult caregiver, and clinician when answering questions (e.g., about the youth’s mental health status). 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 14 for more information on CFIT). 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

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. In order to minimize burden while testing, the battery was divided into two booklets of different (not “repeated”) measures and administered at two consecutive sessions (on average, one week apart). Certain measures such as the Therapeutic Alliance Quality Scale (TAQS) were purposefully placed in the 2nd booklet to help assure that even relatively new clients (and their adult caregiver) would have at least one or two sessions with their clinician before answering questions about the therapeutic relationship. The SFSS and the external measure used to test its validity were placed in the 1st booklet. 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. Offices were asked to administer both booklets to all eligible clients within four weeks then ship their completed materials to Vanderbilt. This time frame was optimistic; data were received from the 1st region about seven weeks after the study’s start date.
A multi-step process began once data were received. First, the number of booklets received was logged by region, respondent type (youth, adult caregiver, or clinician) and administration (1st or 2nd). Second, a detailed protocol was used to check for data quality, including problems of respondents recording two answers for the same item or highly suspect response patterns that would suggest invalid data. Coin toss was used to determine which among two answers for the same item to code, only when the responses represented adjacent categories and were not contradictory. Heads always indicated coding the higher value; tails, the lower. For example, if strongly agree (code 5) and agree (code 4) were marked and the coin toss showed heads, strongly agree (code 5) would be coded for data entry. If agree and disagree were both endorsed, the item was coded as missing. Otherwise, data that remained ambiguous were considered missing. Unusual response patterns, for example, answering all questions identically, were reviewed independently by two raters. There were remarkably few instances of aberrant response patterns, and raters nearly always agreed when one presented itself. The project’s data manager made the final determination in the event of inter-rater disagreement. Cases with any response pattern were flagged, by measure, so that they could be excluded as needed during analysis. Additional safeguards were used for data entry. After initial cleaning, data were entered twice into a Microsoft Office Access database. Special programming alerted data entry staff to discrepancies between the two entries as well as entry of out-of-range values for each variable. The database was translated into SAS system files. Univariate statistics (e.g., frequencies; means) for each variable were generated and examined for accuracy.

As Table 2.1 shows, 610 youths, 486 caregivers, and 603 clinicians completed the 1st booklet. Somewhat fewer completed the 2nd booklet, with completion rates of 85%, 84%, and 88%, respectively.

<table>
<thead>
<tr>
<th>Completed 1st and 2nd Booklets</th>
<th>1st Booklet (#)</th>
<th>2nd Booklet (#)</th>
<th>2nd Booklet Completion (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Youths</td>
<td>610</td>
<td>519</td>
<td>85%</td>
</tr>
<tr>
<td>Adult Caregivers</td>
<td>486</td>
<td>408</td>
<td>84%</td>
</tr>
<tr>
<td>Clinicians</td>
<td>603</td>
<td>532</td>
<td>88%</td>
</tr>
</tbody>
</table>

*Sample Characteristics*

Of the 610 youths who completed the 1st booklet, 42% were female. Over a third (36%) were African American; 56% white; and 8% multi-ethnic or of other ethnic backgrounds. Eighteen percent was Hispanic. On average, youths were 15 years old, with 15% between 11-12 years of age; 45% between 13-15 years of age, and 40% between 16-18 years of age. Based on caregiver reports, over two-thirds (67%) of youths had been diagnosed with a mental health disorder at some time before or with current treatment. Almost half (48%) of the youths had used alcohol or drugs.
On average, caregivers were 45 years old. Nearly 90% were female. About 40% were African American; 57% white; and 3% multi-ethnic or of other ethnic backgrounds. Eleven percent was Hispanic. About 20% of caregivers had not received a high school diploma; two-thirds had completed high school; and 15% had earned a bachelor’s or higher degree. Almost half (45%) had annual household incomes less than $21,000; 37% had incomes of $31,000 or more. About a third of caregivers were divorced or separated, about half were currently married or living as married, 14% were never married, and others were widowed. Two-thirds reported ever being diagnosed with an emotional, behavioral, or substance use problem, but only 20% reported ever receiving professional help.

Nearly all (95%) caregivers considered themselves to be the youth’s primary caregiver. About half were the youth’s birth parent, another 33% were a foster parent, with others as grandparents or other family members. Three-fourths reported they knew the youth at least fairly 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 was determined using several standard methods in addition to methods specifically designed for developing brief instruments for frequent use. These methods are described below. Note that established psychometrics for all types of reliability described below are not available for all instruments (see chapters 4-13 for a discussion of tests for each measure).

Reliability Coefficients
Cronbach’s alpha internal consistency reliability correlations were calculated for total scores and any subscale scores where applicable. This statistic is larger when internal consistency is high, and smaller when it is low. A Cronbach’s alpha of 0.80 or higher is generally considered satisfactory (Clark & Watson, 1995), indicating 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 valid 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 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. Traditionally, psychometricians would say that items with extreme means and kurtosis have floor or ceiling artifacts that make little psychometric contribution to a test.

- **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. Please note that two different types of scaling were used in this manual. Most chapters scaled the Rasch measure scores to have a mean of 50 and standard deviation of 10. Three
chapters (8, 9, and 10) used a scaling with a mean of 0 and a standard deviation of 1.

- **Item-Total Correlation**: In classical test theory (Lord & Novick, 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 contribute to the reliability of measurement and items outside that range do not.

- **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.

We dropped items 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 2.62 would mean that the true severity level for the youth is between 54.76 and 65.24 with approximately 95% certainty. Thus, the smaller the standard error, the more precise the measurement.

**Reliable Change Index**

The reliable change index (RCI) 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 5.36 points (which corresponds to the reliable 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 reliable change threshold. If the change in scores is less than the reliable change 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 reliability over time. This form of reliability is only available for the SFSS at this time. Future updates of the PTPB will include additional analyses for the other instruments. 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 two-week SFSS in their 1st and 2nd booklets and a few items on demographic background and the number of sessions between administrations. Clinicians were instructed to administer the two booklets 4-7 days apart at the end of each of two sessions. If the two sessions were more than 4-7 days apart, they were asked to administer the 1st booklet, then the 2nd at the very next session. Participants for whom there was less than one day or more than 14 days 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 |
|-------------|----------------|----------------|----------------|
| SFSS        | # of Regions   | # of Youth Reports | # of Adult Caregiver Reports | # of Clinician Reports |
|             | 1<sup>st</sup> Adm. | 2<sup>nd</sup> Adm. | 1<sup>st</sup> Adm. | 2<sup>nd</sup> Adm. | 1<sup>st</sup> Adm. | 2<sup>nd</sup> Adm. |
| Reliability Test |                   |                   |                   |                   |
| Total for Reliability Test | 4 | 105 | 84 | 85 | 54 | 108 | 88 |

3 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.
Determining Validity

**Factorial Validity**
One of the chief purposes of the study was to assess the factorial validity of the PTPB measures. Factorial validity examines whether the correlations among test items fit the theory of what the test purports to measure. For some measures where the factorial structure was more complex and not as clearly determined by theory, we split the psychometric sample randomly into two groups, a test half of the sample and a validity sample. Analyses were then conducted with the test sample until a clear factor structure was determined and the model fit the test data in a satisfactory way. Then, this factorial structure was validated using the validity sample.

**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. Most often, scree plots were 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 (CGSQ), are multifactorial, meaning they have subtests. The multiple factors were developed by inspection of scree plots followed by confirmatory factor analysis on a split sample.

We would like to note that scree plots are tools of exploratory factor analyses. Whereas exploratory factor analysis (EFA) plays an important role in identifying the underlying dimensions of an instrument when there are no a priori theoretical expectations, confirmatory factor analysis (CFA) is most useful for evaluating whether a theoretically meaningful model fits the observed data (Floyd & Widaman, 1995).

**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 Root Mean Square Error of Approximation (RMSEA; 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 RMSEA, a value of 0.05 indicates close fit, 0.08 fair fit, and 0.10 marginal fit.
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>RMSEA</td>
<td>Satisfactory: 0.05 or less</td>
</tr>
<tr>
<td></td>
<td>Fair: 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. Only two comparisons were statistically significant. For the clinician report of the SFSS-33, white youth had somewhat higher scores (mean = 63.7, $SD = 8.9$) than minority youth (mean = 60.8, $SD = 7.9$), $F = 10.96, p < 0.001$. For the adult caregiver version of the TOES, a significant age by race interaction was found, which reflects slightly lower outcome expectations for caregivers of minority youth between the ages of 14 and 15 years than caregivers of white youth. 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 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.

**References**


