Published in final edited form as:
Adm Policy Ment Health. 2012 March ; 39(0): 3–12. doi:10.1007/s10488-012-0404-1.

The Peabody Treatment Progress Battery: History and Methods for Developing a Comprehensive Measurement Battery for Youth Mental Health

Manuel Riemer, Ph.D.,
Wilfrid Laurier University

M. Michele Athay, M.S.,
Vanderbilt University

Leonard Bickman, Ph.D.,
Vanderbilt University

Carolyn Breda, Ph.D.,
Vanderbilt University

Susan Douglas Kelley, Ph.D., and
Vanderbilt University

Ana R. Vides de Andrade, Ph.D.
Vanderbilt University

Abstract

There is increased need for comprehensive, flexible, and evidence-based approaches to measuring the process and outcomes of youth mental health treatment. This paper introduces a special issue dedicated to the Peabody Treatment Progress Battery (PTPB), a battery of measures created to meet this need. The PTPB is an integrated set of brief, reliable, and valid instruments that can be administered efficiently at low cost and can provide systematic feedback for use in treatment planning. It includes eleven measures completed by youth, caregivers, and/or clinicians that assess clinically-relevant constructs such as symptom severity, therapeutic alliance, life satisfaction, motivation for treatment, hope, treatment expectations, caregivers strain, and service satisfaction. This introductory article describes the rationale for the PTPB, its' development and evaluation, detailing the specific analytic approaches utilized by the different papers in the special issue and a description of the study and sample from which the participants were taken.

Keywords

Peabody Treatment Progress Battery; PTPB; measurement battery; youth mental health; psychometrics; treatment outcomes

Correspondence concerning this article should be addressed to Michele Athay, Vanderbilt University, Center for Evaluation and Program Improvement, Peabody #151, 230 Appleton Place, Nashville, TN 37203. Electronic mail may be sent to Michele.Athay@vanderbilt.edu.

Ever since Jane Knitzer published her 1984 report on the status of children's mental health services in the U.S., many attempts have been made to improve clinical services for children and youth. No matter what the approach (e.g., evidence-based treatments, system of care, licensing, measurement feedback systems) the only way to know whether these interventions actually lead to improvements for children served in usual treatment settings is to have sound measurement tools for assessing treatment progress and success. A report sponsored by the federal government of Australia identified some key criteria for such measurement tools: feasibility, comprehensiveness, flexibility, potential for improving clinical effectiveness, and psychometric soundness (Bickman, Nurcombe, Townsend, Belle, Schut, & Karver, 1998). The Peabody Treatment Progress Battery (PTPB; Bickman et al., 2007, 2010) featured in this special issue provides such a comprehensive, flexible, and evidence-based approach to assessing the process and outcomes of mental health services for youths aged 11–18 years.

The PTPB includes eleven clinically relevant measures of key mental health outcomes and clinical processes for most types of clinical services available for youth. As this special issue demonstrates, each measure underwent rigorous psychometric testing with the goal of being scientifically sound while simultaneously being brief and clinically useful. The measures, especially with their repeated use, offer clinicians and others the opportunity for 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 who are not improving as expected. As an integrated set of practical, reliable, and valid instruments, the PTPB can be administered efficiently and at low cost. In this paper, we will first provide a brief history of the PTPB followed by a description of the second edition (v.2) of the battery. Then, we will discuss the procedures and samples used for the study from which the psychometric and analytical data were drawn. Finally, we will explain the analytical procedures and statistical indicators that were used for the psychometric evaluation of all measures included in this special issue. Although each article in this issue can stand alone, this introduction provides more in depth information in order to avoid excessive repetition in the subsequent articles.

A Brief History of the PTPB

Some preliminary steps toward the creation of the PTPB included the development of a few individual measures and a comprehensive review of existing measurement instruments for children and youth mental health. The first step was the development of an early version of the Symptoms and Functioning Severity Scale (SFSS) by Bickman, Lambert, and Summerfelt in 1996. Then, from 1997 to 2000, Bickman worked in Australia on measurement system development. Part of this work was supported by a grant awarded by the Commonwealth that resulted in the monograph, *Consumer Measurement System in Child and Adolescent Mental Health* (Bickman, Nurcombe, Townsend, Belle, Schut, & Karver, 1998), reviewing over 100 instruments. Based on this review the authors concluded that existing measures were not suitable for routine measurement because of either length or insufficient psychometric quality. They recommended the development of a new measurement system to better meet psychometric and practice needs. Research continued on

a measure of adolescent functioning (Karver & Bickman, 2002) and on a measure of therapeutic alliance (Bickman et al., 2004).

The actual development of the PTPB as a comprehensive battery began in April 2004, when Bickman received a NIMH grant to study the effectiveness of a measurement feedback system, which required psychometrically sound measures that could be used in routine clinical practice. This study evaluated Contextualized Feedback Systems or CFS™ (previously called CFIT; Bickman, Riemer, Breda, & Kelley, 2006), a comprehensive online continuous quality improvement system for mental health services (Bickman, Kelley, Breda, De Andrade, & Riemer, 2011). Our collaborating partner for this study was a large national social service agency that delivers home and community based services for youth and families. Over the course of five years, over 28 sites across 10 U.S. states were involved in this study. The design and pilot testing (including cognitive interviews) of the first edition PTPB measures took place from May 2004 to April 2005. The collaborating partner was closely involved in the development and the psychometric evaluation of the PTPB measures. Directors, clinical supervisors, and clinicians all provided feedback on the type of measures that should be included in the battery and on the content and structure of the measures. A comprehensive psychometric study was conducted from May to September of 2005, resulting in some final refinements before the publication of the first PTPB manual, which was made available for free in 2007 (Bickman et al., 2007). The PTPB received very positive reviews from both academic and practice leaders and was licensed to over 990 organizations within the first year.

Since the first version was issued, Bickman and his colleagues have conducted further extensive psychometric studies on another large longitudinal sample of youth receiving home based services. One of the measures used in this battery (the Youth Counseling Impact Scale) was published in the premier measurement journal, *Psychological Assessment* (Riemer & Kearns, 2010). The methodology used to assess the psychometric properties of that measure was used for all the other measures in the battery, thus assuring that the highest standards were followed in the measurement development process. The 2010 edition also included the Session Report Form (SRF), a session-by-session measure capturing the content and topics of clinical sessions (Kelley, Vides de Andrade, Sheffer, & Bickman, 2010). Most measures have been further reduced in length during the second psychometric evaluation. This special issue presents the measures as presented in the second edition of the PTPB (Bickman et al., 2010).

Measurement Battery

In contrast to the typical single instrument, the PTPB is an integrated comprehensive set of measures not available elsewhere in the child and adolescent mental health field. Research clearly indicates that progress in treatment is multidimensional (Bickman, Karver, & Schut, 1997), and that information is needed not only on traditional outcomes such as a reduction in symptoms, but also on the treatment process that mediate such outcomes. More than simply focusing on outcomes, the PTPB uses a common factors approach to the measurement of treatment processes (Karver, Handelsman, Fields, & Bickman, 2005). These common factors, such as therapeutic alliance, are elements not particular to any specific therapy (e.g.,

Cognitive Behavioral Therapy), but common across most therapies (Lambert, 2005), and are seen by many as largely responsible for the benefits of therapy. As common factors are not uniquely linked to specific therapies, they allow for the PTPB to be used with almost any type of treatment or intervention.

The PTPB (2nd ed.) contains 11 instruments measuring both therapeutic processes and outcomes, including positive or strength-oriented outcomes (e.g., hope) and more traditional measures of problems (e.g. symptom severity). The PTPB is intended for use with youth aged 11 to 18 years, in varied service settings and clinical programs, including outpatient care, in-home treatment, and foster care. Intensity of treatment can range from multiple sessions within a week to biweekly treatment. The PTPB has not been systematically tested for use in more restrictive service settings such as residential or inpatient treatment, but has been used successfully in at least one residential treatment facility. All of the measures are currently available in English or Spanish, and are written at a fourth-grade reading level. The instruments can be completed individually by the respondent or, if needed, read aloud to a youth or adult caregiver. Plans are underway to extend the PTPB for younger child and adult populations, as well as to translate it into other languages.

Measures Included in the PTPB (2nd ed.)

A total of eleven measures are included in the PTPB (2nd ed.). Each are described below. The first six measures assess constructs considered treatment outcomes (e.g. symptom severity, life satisfaction) and the last five measures assess constructs that assess the process of treatment (e.g. therapeutic alliance, counseling impact).

Symptoms and Functioning Severity Scale (SFSS)

Completed by the youth, adult caregiver and clinician, the SFSS can be used to measure symptom severity at baseline, regularly throughout treatment, and at discharge. The SFSS is best considered a global measure of severity and is not an instrument that can be used to provide a diagnosis. Items are based on four of the most common mental health disorders for youth: ADHD, conduct/oppositional disorder, depression, and anxiety. There are parallel forms (Short Forms A and B) for the youth, caregiver, and clinician. Scores are reported as a total score, with two subscale scores (internalizing and externalizing). See Athay, Riemer, and Bickman (2012) in this issue for more information on the SFSS.

Brief Multidimensional Students' Life Satisfaction Scale – PTPB Version (BMSLSS-PTPB)

Completed by youth, the BMSLSS-PTPB assesses life satisfaction across five dimensions. This short questionnaire (6 items) can be administered on the same schedule as the SFSS, and yields a total score of life satisfaction assessed across six domains. The PTPB version represents a revised version of the BMSLSS (Seligson, Huebner, & Valois, 2003). See Athay, Kelley, and Dew-Reeves (2012) in this issue for more information on the BMSLSS-PTPB.

Satisfaction with Life Scale (SWLS)

A short instrument, the SWLS (Pavot & Diener, 1993) is completed by adult caregivers to measure their global judgments of life satisfaction. This five item questionnaire yields a total score, and has the same schedule as the caregiver strain measure. The SWLS can be administered at baseline, regularly during concurrent treatment, and at discharge. See Athay (2012) in this issue for more information about the SWLS.

Caregiver Strain Questionnaire–Revised Short Form (CGSQ-SF7)

The CGSQ-SF7 assesses the extent to which caregivers and families experience additional demands, responsibilities, and difficulties resulting from caring for a child with emotional or behavioral disorders. Components of caregiver strain include objective strain (i.e., observable negative consequences of caring for someone with special needs) and subjective strain (i.e., caregivers' feelings associated with the objective strain). A shortened version of the original CGSQ (Brannan, Heflinger, & Bickman, 1997), this reduced seven-item version provides a total score and two subscale scores, objective strain and subjective strain. The CGSQ-SF7 is completed by adult caregivers at baseline, a few times during treatment, and at discharge. See Brannan, Athay, and Vides de Andrade (2012) in the current issue for more information about the CGSQ-SF7.

Children's Hope Scale-PTPB Version (CHS-PTPB)

A self-report assessment of the youth's beliefs in the ability to achieve goals, the CHS-PTPB also registers beliefs about initiating and sustaining movement toward these goals. Adapted from Snyder et al.'s Children's Hope Scale (CHS; 1997) the CHS-PTPB provides an overall score of youth hope, and can be administered at baseline, throughout treatment, and at discharge. See Dew-Reeves, Athay, and Kelley (2012) in this issue for more information on the CHS-PTPB.

Service Satisfaction Scale (SSS)

The SSS provides a general indicator of how well youth and adult caregivers perceive the mental health organization's services. The SSS yields a total score and can be completed during concurrent treatment and at discharge. See Athay & Bickman (2012) in the current issue for more information about the SSS.

Therapeutic Alliance Quality Scale (TAQS) and Therapeutic Alliance Quality Rating (TAQR)

The Therapeutic Alliance Quality Scale (TAQS) for youth measures one of the most studied components of effective therapy, the client's relationship with the clinician. The youth version asks five questions concerning the bond the youth has with the clinician and agreement on goals and tasks on a session by session basis. The TAQS provides a total score. The Therapeutic Alliance Quality Rating (TAQR) has a clinician and caregiver version that includes global items on alliance. These serve to orient the clinician when reviewing the youth and adult caregiver versions of the TAQS/TAQR respectively. It is recommended to administer the TAQS and TAQR frequently throughout treatment so that problems with the therapeutic relationship can be detected early to prevent serious disruptions. See Bickman, Vides de Andrade, Athay, Chen, De Nadai, Jordan-Arthur, and

Karver (2012) in this issue for more information about the TAQS and TAQR. The therapeutic alliance measures also include the ratings that clinicians make of their perceptions of how the youth and caregiver rated the alliance for that session. Thus we are able to compare the clinicians' perceptions with the actual ratings provided by the client and caregiver. Data about this use of therapeutic alliance ratings are not included in the Bickman et al. (2012) article in this issue.

Treatment Outcome Expectations Scale (TOES)

The TOES assess youths' and adult caregivers' expectations about the anticipated outcomes of treatment. Completed by the youth and the adult caregiver, the TOES provides a total score, and is administered at baseline only. It may be accompanied by the Treatment Process Expectations Index (TPEI), an additional list of nine recommended questions that assess youth and caregiver expectations about their role in counseling and the counseling process itself. See Dew-Reeves and Athay (2012) in this issue for more information on the TOES.

Youth Counseling Impact Scale (YCIS)

A self-report questionnaire, the YCIS assesses the youth's judgments of the short-term positive impact of counseling in regard to increased insight as well as positive changes in behavior, cognition or affect following the previous session. The YCIS provides a total score and subscale scores for insight and change. See Riemer and Kearns (2009), and Kearns and Athay (2012; this issue) for more information about the YCIS.

Motivation for Youth's Treatment Scale (MYTS)

The MYTS assesses treatment motivation, a key predictor of seeking and staying in services, as well as of treatment outcomes. There are versions for the youth and adult caregiver. Both provide a total score, with subscale scores for problem recognition and treatment readiness. There is a slightly different version for use at baseline and during the treatment phase. See Breda and Riemer (2012) in this issue for more information about the MYTS.

Session Report Form (SRF)

The SRF is a 25-item self-report measure completed by the clinician at the end of each clinical session intended to capture the session content and topics addressed in each treatment session. The SRF is completed every session during treatment and discharge. See Kelley, Vides de Andrade, Sheffer, & Bickman (2010) and Kelley, Vides de Andrade, Bickman, and Robin (2012; this issue) for more information on the SRF.

Each of these measures were included in the previously mentioned measurement feedback study led by Bickman and funded by NIMH, for which the procedures are described next.

Procedures

Participants for the psychometric evaluation and the substantive analyses presented in this special issue were drawn from a larger study evaluating the effects of CFS™ on youth outcomes (Bickman et al., 2011). From 2006 to 2009, this study collected data from youth, their caregivers, and clinicians across 28 regional offices in 10 different states, which are

part of a large national provider for home-based mental health services, primarily focused on youth. The service provider does not prescribe any specific type of treatment modality but encourages them to follow a strength-based approach. Services could include individual and family in-home counseling, intensive in-home services, crisis intervention, substance abuse treatment, life skills training, and case management. Clinicians report using various therapeutic approaches, including cognitive-behavioral, integrative-eclectic, behavioral, family systems, and play therapy. Data were collected by clinical staff at the end of a clinical session and later entered by administrative staff into CFS™, a web-based measurement feedback system. No researchers were present during data collection but quality controls were regularly conducted. For example, a random selection of completed paper versions were compared to data entered into the system. The data in the system was also regularly monitored by clinical supervisors and directors and the researchers to detect any irregularities and delays in entering the data and to observe the use of the data by the clinicians. For this purpose, a comprehensive quality control dashboard was included in the online measurement feedback system. The de-identified data were downloaded and processed following a rigorous procedure developed by the Centre for Evaluation and Program Improvement (Smith, Breda, Simmons, Lambert, & Bickman, 2009).

Eligibility

Eligible youths were 11–18 years old, 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 anyone was present at the time instruments were completed. 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.

Measure Administration

Each youth entered into the feedback system (CFS™) 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 schedule—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 to help with comprehension, but were instructed not to help with answers. All youth and adult caregiver measures were available in English and Spanish. After 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 then data entered into the feedback system. Once data were entered, questionnaires were scored according to their individual psychometrics and an online feedback report became

available. The database was translated into system files for the statistical software SAS©. Univariate statistics (e.g., frequencies; means) for each variable were generated and examined for accuracy.

Confidentiality

Vanderbilt (i.e. researchers) 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. The study was reviewed and approved by Vanderbilt's Institutional Review Board.

Samples Used

Each article in this special issue highlights one of the measures included in the PTPB (2nd ed.) and most present two separate sets of analyses. First, each article presents the results of a comprehensive psychometric evaluation of the measure in a large sample of clinically-referred youth. The second part of the article presents an application or substantive study with that measure expanding existing knowledge in each area. These two sets of analyses include different, but overlapping samples of youth, caregivers, and clinicians. The application or substantive portion of each article utilizes only participants whom were included in the CFS™ evaluation study mentioned above. This evaluation sample included only youth (and their respective caregivers and clinicians) who started their treatment during the data collection period. Youth who had started their treatment prior to data collection were not included in the substantive analyses. This was done so that youth baseline clinical characteristics can be discussed. However, although the data gathered from participants already in treatment when the evaluation commenced were not included in the evaluation, they were included in the sample used for psychometric evaluation of each measure. For those with more than one data point for a given measure, the first available data point was used for psychometric analyses. Similarities and differences between these two different but related samples can be seen in Table 1. The demographics of the participants are comparable across samples.

Data collection occurred under real world conditions where measures were administered by clinical staff rather than researchers. While this procedure made a longitudinal study of this scope possible, it also had drawbacks, such a relatively high level of missing data. During the data collection period, youth entered and left treatment (and thus the data collection) at different times. Therefore, the composition and size of the sample varies throughout time. Also, although measures were scheduled to be administered at the end of clinical sessions, administration was the responsibility of each clinician. Many clinicians reported that they skipped measure administration when the clinical session was considered a crisis session. Additionally, changes to the measurement schedule were made by researchers in 2008 in order to reduce time burden. This resulted in less frequent data points for some measures from this time point on and for the YCIS and MYTS measures data collection was completely stopped. Given all these factors, the number of participants with completed measures varies by measure and sample. Table 2 shows the number of participants, by sample, with at least one completed measure. Measures were counted as completed if 85%

of the items were answered. Mean imputation was used for cases with less than 100% of the items completed. If more than 15% of the items were missing, the measure score was not computed and was counted as missing.

Psychometric Evaluation

Based on our review of the literature, we identified the following quality criteria as desirable for the PTPB:

- Reliable: Every question in a battery must contribute accurate information.
- Valid: Scores must have evidence-based interpretations.
- Brief: 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.

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 special issue presents the results of a second round of comprehensive psychometric testing on the measures included in the PTPB. For more information concerning the initial development and first round of psychometric testing, see Bickman and colleagues (2007).

The psychometric analyses emphasized evaluating every item of every measure for its reliability and validity. Multiple methods were used including those from classical test theory (CTT), 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. We inspected reliability, comprehensive item psychometrics, and validity and compared results with known standards. Minimum detectable change indices are also calculated for measure scores. Methods and standards are described below.

Reliability

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).

The Standard Error of Measurement (*SEM*) shows how much uncertainty there is around each youth's score on a given occasion. It can be estimated by the formula $SEM = SD[(1 - r)^{1/2}]$, where *SD* is the baseline standard deviation and *r* is the test–retest reliability

coefficient. 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). Thus, the smaller the standard error, the more precise the measurement. Because we did not generate test-retest samples for the measures in the PTPB, we used the internal reliability coefficient alpha as an approximation.

Minimum Detectable Change

Based on the *SEM* values for the minimum detectable change (MDC) was calculated. The MDC represents the smallest change in scores from one measurement instance to the next that likely reflects true change rather than chance and measurement error alone (Schmitt & Di Fabio, 2004). The level of certainty represented by the MDC is determined by the respective Z-score that is used in calculating it. The MDC is calculated by multiplying the *SEM* by the z-score associated with the desired level of confidence and the square root of 2. For practical purposes of clinical decision making it was decided to set the confidence level at 75%. This means, if a youth's self report score on the youth version of the SFSS decreased by more than 4.63 points (which corresponds to the MDC reported for the Youth SFSS), for example, one 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 MDC. If the change in scores is less than the MDC, the clinician may want to determine in some other way if a change in scores represents meaningful 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.

In addition to the MDC, 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 according to quartiles.

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 issue, 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 (Bond & Fox, 2001). Table 3 shows that statistical information collected for the items in each

measure and the criteria used to evaluate them. The use of IRT modeling may not be as familiar as CTT methods. Therefore, we will expand on the IRT criteria listed in Table 3.

- 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.
- 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 Wright and Linacre (1994), items with an *Infit* and *Outfit* between 0.6 and 1.4 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) provides estimates of each item's discrimination (a second item parameter) after the 1-parameter logistic model is estimated. Items with low discrimination (i.e., significantly less than 1) are less effective at differentiating between people who are high or low on the measured trait.

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. The current psychometric evaluations presented in this issue represent a second round of analyses with an independent sample. Thus, the current issue aims to confirm the structure of each measure as found in the first edition of the PTPB. Therefore, confirmatory factor analysis (CFA) is used. 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. 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). In most cases we compared multiple models (e.g., a two-factor model vs. a one-factor model) to assess that the existing model is indeed the one fitting the data the best.

Two other forms of validity were assessed during the first psychometric evaluation of the PTPB: convergent and discriminant (or divergent) validity (Campbell & Fiske, 1959). Convergent validity requires that measures of similar constructs should be positively correlated. For example, we validated the Symptoms and Functioning Severity Scale (SFSS) score by determining how it correlated with similar measures, namely the Child Behavior Checklist (CBCL; Achenbach, 1991), the Youth Self Report (YSR; Achenbach), the Youth Outcomes Questionnaire (Y-OQ®; Wells, Burlingame & Lambert, 1999), and the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1999). The correlations, approximately r = 0.80, suggested that the SFSS was very similar to the other instruments. They are all measures of reported emotional and behavioral problems (see Bickman et al., 2007 for more details). Given high correlations (r = .99; see Table 4) between the first edition's SFSS-33 total scores and the current edition's SFSS-Full total scores for all respondents (youth, clinician and adult caregiver), we determined it unnecessary to repeat the extensive data collection required to replicate our previous convergent validation analyses.

To assess discriminant validity we first correlated the adult caregiver and clinician ratings of youth symptoms and functioning (SFSS-Full) with three treatment process measures (SSS, MYTS, and TOES). If there is discriminant validity within this battery, these measures should measure different constructs resulting in small bivariate correlations. Overall, the correlations are small (r _< 0.30) indicating good discriminant validity. The only exception is the correlation between the caregiver SFSS scores and the caregiver MYTS ratings (r=0.57). However, this relationship was expected because the MYTS includes a problem recognition subscale that should be directly related to the caregiver's perception of symptom severity. We then correlated three youth-rated outcome measures (SFSS, CHS-PTPB, and BMLSS-PTPB) with five youth-rated process measures (YCIS, SSS, TAQS, TOES, and MYTS). Life satisfaction and hope ratings had small correlations with process scores. Symptom scores, on the other hand, were not correlated with process measures except for the expected positive correlation with treatment motivation (MYTS) and therapeutic alliance (TAQS). The correlation with MYTS was expected because higher treatment motivation usually accompanies serious symptoms. The small negative correlation with TAQS does not pose a threat to discriminant validity. More details for these analyses can be found in the second edition of the PTPB manual (Bickman, et al., 2010).

This issue represents the effort of many to develop a battery of measures that could be used concurrently with treatment of youth receiving mental health services. The battery is multidimensional as represented by the divergent validity described above. The measures have strong psychometric validity when used with this population. As the articles in this special issue demonstrate they can be used productively for research purposes to discover new information about youth receiving mental health services as well what treatments might work best with them. Of prime importance they are practical measures that can be used in real world settings. Combined with a measurement feedback system such as CFS they provide an effective way to not only measure what is happening to youth receiving treatment but also provide an approach to improving outcomes along many dimensions. The measures are licensed by Vanderbilt University but are available at no cost in the paper and pencil version. Only a completed registration is required (http://peabody.vanderbilt.edu/ptpb).

Acknowledgments

This research was supported by NIMH grants R01-MH068589 and 4264600201 awarded to Leonard Bickman. We would like to thank the clinicians, youth and caregivers for providing us with these data and the Providence Service Corporation for their enthusiastic participation over several years to gather this information. Finally, thanks to Sarah (Sally) Horwitz for serving as action editor for this volume.

References

Achenbach, TM. Integrative guide for the 1991 CBCL/4–18, YSR, and TRF profiles. Burlington, VT: University of Vermont, Department of Psychiatry; 1991.

Athay MM. The Satisfaction with Life Scale (SWLS) in caregivers of clinically-referred youth: Psychometric properties and mediation analysis. Administration and Policy in Mental Health and Mental Health Services Research. 2012

Athay MM, Bickman L. Development and psychometric evaluation of the youth and caregiver Service Satisfaction Scale. Administration and Policy in Mental Health and Mental Health Services Research. 2012 doi:

Athay MM, Kelley SD, Dew-Reeves SE. Brief Multidimensional Students' Life Satisfaction Scale – PTPB Version (BMSLSS-PTPB): Psychometric properties and relationship with mental health symptoms over time. Administration and Policy in Mental Health and Mental Health Services Research. 2012

Athay MM, Riemer M, Bickman L. The Symptoms and Functioning Severity Scale (SFSS): Psychometric evaluation and discrepancies among youth, caregiver, and clinician ratings over time. Administration and Policy in Mental Health and Mental Health Services Research. 2012

Bentler PM. Comparative fix indexes in structural models. Psychological Bulletin. 1990; 107(2):238–246. [PubMed: 2320703]

Bickman L, Kelley S, Breda C, De Andrade A, Riemer M. Effects of routine feedback to clinicians on youth mental health outcomes: A randomized cluster design. Psychiatric Services. 2011; 62(12): 1423–1429. [PubMed: 22193788]

Bickman, L.; Athay, MM.; Riemer, M.; Lambert, EW.; Kelley, SD.; Breda, C.; Tempesti, T.; Dew-Reeves, SE.; Brannan, AM.; Vides de Andrade, AR., editors. Manual of the Peabody Treatment Progress Battery, 2nd ed. [Electronic version]. Nashville, TN: Vanderbilt University; 2010. http://peabody.vanderbilt.edu/ptpb\

Bickman L, Karver M, Schut LJA. Clinician reliability and accuracy in judging appropriate level of care. Journal of Consulting and Clinical Psychology. 1997; 65(3):515–520. [PubMed: 9170776]

Bickman, L.; Nurcombe, B.; Townsend, C.; Belle, M.; Schut, J.; Karver, M. Consumer measurement systems for child and adolescent mental health. Canberra, ACT: Department of Health and Family Services; 1998. http://www.health.gov.au/hsdd/mentalhe

Bickman L, Riemer M, Breda C, Kelley SD. CFIT: A system to provide a continuous quality improvement infrastructure through organizational responsiveness, measurement, training, and feedback. Report on Emotional and Behavioral Disorders in Youth. 2006; 6:86–87. 93–94.

Bickman, L.; Riemer, M.; Lambert, EW.; Kelley, SD.; Breda, C.; Dew, SE.; Brannan, AM.; Vides de Andrade, AR., editors. Manual of the Peabody Treatment Progress Battery [Electronic version]. Nashville, TN: Vanderbilt University; 2007. http://peabody.vanderbilt.edu/ptpb\

Bickman L, Vides de Andrade AR, Athay MM, Chen JI, De Nadai AS, Jordan-Arthur B, Karver MS. The role of therapeutic alliance in predicting improvement in youth outcomes: Whose ratings matter the most? Administration and Policy in Mental Health and Mental Health Services Research. 2012

Bickman L, Vides de Andrade AR, Lambert EW, Doucette A, Sapyta J, Boyd AS, et al. Youth therapeutic alliance in intensive treatment settings. Journal of Behavioral Health Services and Research. 2004; 31(2):134–148. [PubMed: 15255222]

Breda CS, Riemer M. Motivation for Youth's Treatment Scale (MYTS): A new tool for measuring motivation among youths and their caregivers. Administration and Policy in Mental Health and Mental Health Services Research. 2012 doi:

Bond, TG.; Fox, CM. Applying the Rasch model : fundamental measurement in the human sciences. Mahwah, NJ: L. Erlbaum; 2001.

Brannan AM, Athay MM, Vides de Andrade AR. Measurment qualitly of the Caregiver Strain Questionnaire – Short Form 7 (CGSQ-SF7). Administration and Policy in Mental Health and Mental Health Services Research. 2012 doi:

Brannan AM, Heflinger CA, Bickman L. The caregiver strain questionnaire: Measuring the impact on the family of living with a child with serious emotional disturbance. Journal of Emotional and Behavioral Disorders. 1997; 5(4):212–222.

Brown W. Some experimental results in the correlation of mental abilities. British Journal of Psychology. 1910; 3:296–322.

Browne, MW.; Cudeck, R. Alternative ways of accessing model fit. In: Bollen, KA.; Long, JS., editors. Testing structural equation models. Newbury Park: Sage; 1993. p. 136-162.

Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin. 1959; 56(2):81–105. [PubMed: 13634291]

Cronbach LJ, Shavelson RJ. My Current Thoughts on Coefficient Alpha and Successor Procedures. Educational and Psychological Measurement. 2004; 64(3):391–418.

Dew-Reeves SE, Athay MM. Validation and Use of the Youth and Caregiver Treatment Outcome Expectation Scale (TOES) to Assess the Relationships between Expectations, Pretreatment Characteristics, and Outcomes. Administration and Policy in Mental Health and Mental Health Services Research. 2012 doi:

Dew-Reeves SE, Athay MM, Kelley SD. Validation and use of the Children's Hope Scale – PTPB Edition (CHS-PTPB): High initial youth hope and elevated baseline symptomology predict poor treatmetn outcomes. Administration and Policy in Mental Health and Mental Health Services Research. 2012 doi:

Goodman R. The extended version of the Strengths and Difficulties Questionnaire as a guide to child psychiatric caseness and consequent burden. The Journal of Child Psychology and Psychiatry and Allied Disciplines. 1999; 40(5):791–799.

Harlow, LL. The Essence of Multivariate Thinking. Mahwah, NJ: Lawrence Erlbaum Associates, Inc; 2005.

Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling. 1999; 6(1):1–55.

Joreskog, KG. Analysis of covariance structures. In: Nesselroade, JR.; Cattell, RB., editors. Handbook of multivariate experimental psychology. 2nd ed.. New York, NY: Plenum Press; 1988. p. 207-230.

Karver MS, Bickman L. Positive functioning: Does it add validity to maladaptive functioning items? Evaluation and Program Planning. 2002; 25(1):85–93.

Karver MS, Handelsman JB, Fields S, Bickman L. A theoretical model of common process factors in youth and family therapy. Mental Health Services Research. 2005; 7(1):35–51. [PubMed: 15832692]

Kearns MA, Athay MM. Measuring youths' perceptions of counseling impact: Description, psychometric evaluation, and longitudinal examination of a revised Youth Counseling Impact Scale. Administration and Policy in Mental Health and Mental Health Services Research. 2012 doi:

Kelley SD, Vides de Andrade AR, Bickman L, Robin A. The Session Report Form (SRF): Are clinicians addressing issues of concern to youth and caregivers? Administration and Policy in Mental Health and Mental Health Services Research. 2012 doi:

Kelley SD, Vides de Andrade AR, Sheffer E, Bickman L. Exploring the black box: Measuring youth treatment process and progress in usual care. Administration and Policy in Mental Health and Mental Health Services Research. 2010; 37(3):287–300. doi:10.1007/s10488-010-0298-8. [PubMed: 20238155]

Knitzer J. Mental health services to children and adolescents: A national view of public policies. American Psychologist. 1984; 39:905–911. [PubMed: 6476584]

Lambert MJ. Special section on the placebo concept in psychotherapy. Early response in psychotherapy: Further evidence for the importance of common factors rather than “placebo effects.”. Journal of Clinical Psychology. 2005; 61(7):855–869. [PubMed: 15827996]

Linacre, JM. WINSTEPS® 3.63.0.[Computer software]. 2007. Retrieved Jan 8, 2007, from http://www.winsteps.com/index.htm.

Linacre, JM.; Wright, BD. WINSTEPS Rasch-model computer program (Version 3.47) [computer software]. Chicago: MESA Press; 2006.

Nunnally, JC.; Bernstein, IH. Psychometric theory. 3rd ed.. New York: McGraw-Hill; 1994.

Pavot W, Diener E. Review of the Satisfaction with Life Scale. Psychological Assessment. 1993; 5(2): 164–172.

Riemer M, Kearns MA. Description and psychometric evaluation of the youth counseling impact scale. Psychological Assessment. 2010; 22(10):259–268. [PubMed: 20528053]

Seligson J, Huebner ES, Valois RF. Preliminary validation of the Brief Multidimensional Students' Life Satisfaction Scale (BMSLSS). Social Indicators Research. 2003; 61:121–145.

Smith, CM.; Breda, CB.; Simmons, TM.; Lambert, EW.; Bickman, L. Data preparation and data standards: The devil is in the details. In: Stiffman, AR., editor. The field research survival guide. New York: Oxford University; 2009. p. 82-103.

Schmitt JS, Di Fabio RP. Reliable change and minimum important difference (MID) proportions facilitated group responsiveness comparisons using individual threshold criteria. Journal of Clinical Epidemiology. 2004; 57:1008–1018. [PubMed: 15528051]

Snyder CR, Hoza B, Pelham WE, Rapoff M, Ware L, Danovsky M, et al. The development and validation of the Children’s Hope Scale. Journal of Pediatric Psychology. 1997; 22(3):399–421. [PubMed: 9212556]

Spearman C. Correlation calculated from faulty data. British Journal of Psychology. 1910; 3:171–195.

Steiger JH. Point estimation, hypothesis testing, and interval estimation using the RMSEA: Some comments and a reply to Hayduck and Glaser. Structural Equation Modeling: A Multidisciplinary Journal. 2000; 72(2):149–162.

Wells, MG.; Burlingame, GM.; Lambert, MJ. Youth Outcome Questionnaire. In: Maruish, ME., editor. The use of psychological testing for treatment planning and outcome assessment. 2nd ed.. Mahwah NJ: Lawrence Erlbaum; 1999.

Wright BD, Linacre JM. Reasonable mean-square fit values. Rasch Measurement Transactions. 1994; 8:370. Retrieved March 10, 2011 from www.rasch.org/rmt/rmt83b.htm.

Yu, C-Y. Doctoral dissertation. Los Angeles: University of California; 2002. Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes. 2002 (UMI ProQuest ATT 3066425).

Table 1

Descriptives by Sample

	Psychometric Sample	Evaluation Sample
Youth	N=809	N=356
Gender	46% Female	50% Female
Age	Mean = 14.7 years (*SD*=1.9)	Mean= 15 (*SD*= 1.8)
	49% Caucasian	55% Caucasian
Race	30% African American	25% African American
	17% Hispanic	12% Hispanic
Caregivers	N = 695	N=431
Gender	86% Female	86% Female
Age	Mean = 44.5 years (*SD* =10.6)	Mean = 43 years (*SD*= 10.2)
	57% Caucasian	65% Caucasian
Race	33% African American	27% African American
	12% Hispanic	7% Hispanic
Marriage		
	46% Married or living as married	42%Married or living as married
Status		
	44% Less than $20,000	47% Less than $20,000
Yearly		
	26% $20,000-$34,999	25% $20,000-$34,999
Household		
	16% $35,000 - $49,999	15% $35,000 - $49,999
Income		
	15% $50,000 or more	13% $50,000 or more
	24% less than High School	24% less than High School
Highest		
	60% High School diploma	60% High School diploma
Education		
	16% bachelors or higher degree	16% bachelors or higher degree
Clinician	N=301	N=167
Gender	78% Female	78% Female
Age	Mean = 38 years (*SD* = 11.4)	Mean = 37 years (*SD* = 10.6)
	61% Caucasian	65% Caucasian
Race	28% African American	33% African American
	10% Hispanic	6% Hispanic
Education	80% Masters degree	72% Masters degree

Table 2

Total N's, by Measure and Sample

Measure	Psychometric Sample N	Evaluation Sample N	Measure	Psychometric Sample N	Evaluation Sample N
Youth SFSS	760	340	Clinician TAQR	641	
Caregiver SFSS	686	307	YCIS	462	150
Clinician SFSS	710	294	Caregiver MYTS	457	174
BMSLSS-PTPB	694	334	Youth MYTS	504	197
CHS-PTPB	521	243	SWLS	610	288
Caregiver TOES	268	146	CGSQ-SF7	442	208
Youth TOES	291	175	Caregiver SSS	383	135
Youth TAQ	679	288	Youth SSS	490	195
Caregiver TAQ [1]	561	225			

[1] Caregiver TAQ represented here from the first edition of the PTPB. Data collected on the caregiver TAQR was from a separate study. More details can be found in Bickman, Videsde Andrade, etal., (2012) in this issue. Note: SRF is a session-based report form that requires different analyses and is not represented here. See Bickman et al., 2010 for more information on the SRF.

Table 3

Statistical Properties of Effective Test Items

Criterion	Sought Values	Rationale
Mean	Between 2 and 4 (5 pt scale)	Avoid floors and ceilings in the target sample
Rasch Measure Score (IRT)	Cover the range	Need items across the range of youth
Kurtosis	Less than 2.0[1]	Avoid items where everyone gives same response
Item-Total Correlation.	Higher better	Keep items that measure a single construct
Infit & Outfit (IRT)	Between 0.6 and 1.4[2]	Keep items that fit 1PL (logistic) model
Discrimination (IRT)	Avoid low discrimination	Avoid items that can't discriminate

Note: IRT with WINSTEPS 3.63.(Lincare, 2007)

[1] Harlow (2005)

[2] Wright and Linacre (1994)

Table 4

Correlations of SFSS forms across PTPB editions

	Correlation with SFSS-33 (v.1)	N
Adult Caregiver SFSS-Full (v.2)	0.99	647
Clinician SFSS-Full (v.2)	0.99	671
Youth SFSS-Full (v.2)	0.99	722