# Factor Structure of the Treatment Outcome Package for Children

▼

David R. Kraus
*Behavioral Health Laboratories, Inc.*

▼

James F. Boswell, Aidan G.C. Wright, Louis G. Castonguay, and Aaron L. Pincus
*The Pennsylvania State University*

The Treatment Outcome Package (TOP) is a behavioral health assessment and outcome battery with modules that assess a wide array of behavioral health symptoms and functioning, demographics, case-mix variables, and treatment satisfaction. The TOP was developed to follow the design specifications set forth by the Core Battery Conference (Horowitz, Lambert, & Strupp, 1997), but also to ensure the battery's applicability to naturalistic treatment. The TOP also includes a child version that addresses recommendations recently voiced by several experts in youth assessment and treatment. The purpose of this manuscript is to document the process involved in developing the shortened version of the Child TOP. With large samples of diverse patients, 103 items were reduced to 48, with 13 stable and clinically useful subscales. The final confirmatory factor analysis of 7,267 patients replicated the model with excellent results. Although some of these subscales were similar to factors derived from the adult version of the TOP, others were specific to childhood disorders. In addition, the analyses demonstrated that the TOP is not restricted to manifestations of distress and impairment, but also captures a unique factor of childhood strengths (i.e., resiliency). © 2010 Wiley Periodicals, Inc. J Clin Psychol 66:627–640, 2010.

Keywords: treatment outcomes; assessment tests; test construction

The Treatment Outcome Package (TOP) was designed to measure outcomes in naturalistic settings and to meet the requirements of a core outcome battery as defined by the Society for Psychotherapy Research (SPR) and American

Correspondence concerning this article should be addressed to: David R. Kraus, 293 Boston Post Road, West, Marlborough, MA 01752; e-mail: dkraus@bhealthlabs.com

JOURNAL OF CLINICAL PSYCHOLOGY, Vol. 66(6), 627–640 (2010)

Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/jclp.20675

Psychological Association's (APA) Core Battery Conference (Horowitz, Lambert, & Strupp, 1997). As recently shown by Kraus, Seligman, and Jordan (2005), the adult version of the TOP has demonstrated very good psychometric qualities in a wide range of treatment settings and clinical populations. However, because it cannot be assumed that an instrument developed to measure adult treatment outcome would be valid at all ages, a child version of the TOP was also constructed, introducing items typically unique to children. This article presents the initial development of the third version of the Child TOP, with the primary goal of reducing scale length and increasing model fit statistics.

As argued by Kazdin (2005), the field has been neglectful in providing useful assessment methods for youth clinical work. Not only does the assessment of child psychopathology have a shorter history than the assessment of child abilities and achievement (Achenbach, 2005), but many of the commonly used assessment methods are not based on solid empirical evidence (Mash & Hunsley, 2005). As cogently pointed out by Mash and Hunsley (2005) "gaps exist between the applicability and use of assessments developed and used in research versus real-world settings" (p. 365). Although the Child TOP was developed to be used in naturalistic settings, it also meets numerous recommendations that have been voiced by experts with regards to the development of adequate measures of child psychopathology. Specifically, as recommended by Achenbach (2005) and Kazdin (2005), the Child TOP provides a broad assessment of symptoms and disorders, which is important considering the level of comorbidity as well as the multiple targets frequently aimed at by treatment. Also in line with Achenbach (2005) and Kazdin's (2005) recommendations, rather than being limited to the assessment of psychopathology, the Child TOP also captures adaptive functioning that should be considered by clinicians in their case formulation and evaluation of treatment impact.

The Child TOP monitors patient progress during the course of therapy, as opposed to being restricted to pre–post outcome assessments—a recommendation that has again been voiced by more than one expert in the field of youth treatment (Achenbach, 2005; Kazdin, 2005; Mash & Hunsley, 2005). Constructed to be used in day-to-day practice, the Child TOP meets a large number of specific characteristics identified by Kazdin (2005) as being needed for the clinical application of treatment evaluation. It was developed with the goals of being acceptable to patients and therapists, sufficiently brief and user friendly to be administered within the constraints of clinical practice, and useful across different treatments and clinical problems. In addition, the Child TOP graphically displays positive and negative scale scores that have real-life references (i.e., standard deviations above or below normative data), as well as both positive and negative significant change (also in terms of standard deviations) across assessment points. When comparing the naturalistic outcomes of one clinician or service against another, raw Child TOP subscale change scores can be risk-adjusted to take into account the influence of a number of potentially important case-mix variables (e.g., demographic variables and medical history; Hermann, 2003). With the assessment of case-mix variables that help to further contextualize a client's profile, the interpretations derived from the scale scores are likely to be more useful and the opportunity to study process-related variables in research contexts is facilitated.

Considering the purpose for which it was constructed, as well as the numerous features and characteristics mentioned above, the Child TOP seems particularly relevant to address the paucity of clinically adequate and useful tools to measure child treatment outcomes. The goal of the current study was to use large samples to

further reduce the number of Child TOP items into clinically meaningful subscales and begin assessing its psychometric properties via exploratory and confirmatory factor analyses (Brown, 2006; Floyd & Widaman, 1995; Wiggins, 1973).

## The Treatment Outcome Package for Children

Initial development of the Child TOP scales used a deductive/rational item development strategy (Burisch, 1986) by which the first author (DK) generated more than 250 atheoretical items that spanned diagnostic symptoms and functional areas identified in the Diagnostic and Statistical Manual of Mental Disorders, fourth edition (DSM-IV; American Psychiatric Association, 1994). All DSM-IV Axis I diagnostic symptoms were reviewed, and those symptoms that the first author thought clients could reliably rate on a self-report measure were formulated into questions. Many alternative assessment tools were also reviewed for item inclusion, but most were judged to be based on theoretical constructs inconsistent with DSM-IV symptomatology.

The initial item pool was then presented to other clinicians for edit and review. These clinicians made suggestions for modifications and deletions, based on relative importance and clarity of the items. Parents of clients receiving psychiatric/psychological treatment completed the initial versions of the questionnaires and were asked to provide feedback as well. Questions were reworded based on feedback, and items that were less important or appeared to measure a similar symptom were eliminated. Using principal component analysis (PCA) and increasing sample sizes for each of the iterations, the tool was iteratively revised and reintroduced for feedback. The instrument evaluated here was the result of three iterations of that process.

The current version of the Child TOP is a battery of distinct modules that can be administered all together or in combinations as needed. Expert clinical and client review in the development process ensured adequate face validity. The various modules of the Child TOP include: chief complaints, demographics, treatment utilization and provider characteristics, co-morbid medical conditions and medical utilization, assessment of life stress, treatment satisfaction, functioning, and mental health.

The present study focused on reducing the 103 items of the second version of the Child TOP designed to measure functioning and mental health symptoms. In this article, we describe the development of the 48-item questionnaire and its internal structure, using both exploratory (PCA) and confirmatory methods.

## Method

The 103 mental health symptom and functioning items from the Child TOP administered to a large sample of parents of newly admitted psychiatric clients were analyzed. Participants were instructed to rate each question in relation to "How much of the time during the last two weeks has the child …?" All questions were answered on a 6-point Likert scale: 1 (*all*), 2 (*most*), 3 (*a lot*), 4 (*some*), 5 (*a little*), 6 (*none*).

### *Participants*

The sample comprised 14,030 patients, treated in 299 different behavioral health centers across the United States, whose parent (nonparents, e.g., children who completed the questionnaire themselves, were eliminated from this analysis) completed the child version of the TOP at intake as part of standard treatment. For the participants who provided specific demographic information, the mean age

for the total sample was 11.02 years ($SD = 3.50$). Females comprised 39.7% of the sample. The majority of the sample identified as white/Caucasian (70.2%), followed by African American (13.5%), Hispanic (6.6%), Native American (2.6%), and Asian (0.3%). Approximately 6.8% chose to identify themselves as "other." For those who provided family income, 30.5% reported an income between 0 and $10,000, 25.9% reported between 10 and $20,000, 15.7% reported between 20 and $30,000, 10.1% reported between 30 and $40,000, 6.9% reported between 50 and $75,000, 6.4% reported between 40 and $50,000, 2.9% reported between 75 and $100,000, and 1.6% reported over $100,000. Specific demographic information for the three subsamples is presented in Table 1. A breakdown of service facility types is presented in Table 2.

### *Procedure*

Subjects with missing data were eliminated from an initial sample of 14,030 to simplify confirmatory factor analysis (CFA) computations. The remaining sample of 12,165 subjects was split into three random subsamples to facilitate cross-validation analyses. Samples 1 and 2 ($n$'s = 2,449 each) were used to develop and refine a factor model that was subsequently confirmed in sample 3 ($n = 7,267$).

We conducted missing item analyses to determine if the missing data were missing systematically or if dropping these participants from subsequent analyses would

Table 1
*Demographic Information for the Study Subsamples*

| | Age | | Gender (%) | |
|---|---|---|---|---|
| | *M* | *SD* | Male | Female |
| Sample 1 | 11.05 | 3.52 | 59.7 | 40.3 |
| Sample 2 | 11.08 | 3.49 | 60.5 | 39.5 |
| Sample 3 | 11.00 | 3.50 | 60.5 | 39.5 |
| | | Race/ethnicity (%) | | |
| | Sample 1 | Sample 2 | Sample 3 | |
| White/Caucasian | 70.4 | 70.9 | 69.9 | |
| African American | 13.0 | 13.6 | 13.6 | |
| Hispanic | 6.6 | 6.1 | 6.9 | |
| Native American | 3.0 | 2.6 | 2.5 | |
| Asian | 0.2 | 0.1 | 0.4 | |
| | | Income (%) | | |
| | Sample 1 | Sample 2 | Sample 3 | |
| 0–10 | 28.8 | 31.3 | 30.8 | |
| 10–20 | 26.2 | 26.0 | 25.9 | |
| 20–30 | 16.3 | 15.3 | 15.6 | |
| 30–40 | 9.8 | 9.7 | 10.2 | |
| 40–50 | 6.8 | 6.1 | 6.4 | |
| 50–75 | 8.1 | 6.9 | 6.6 | |
| 75–100 | 2.8 | 2.7 | 3.0 | |
| Over 100 | 1.2 | 2.0 | 1.5 | |

*Note*. M, Mean; SD, standard deviation. Percentages (%) based on participants who provided a response. Income reported in thousands of dollars. Sample comparisons using ANOVA and $\chi^2$ tests indicated no significant differences between the samples.

Table 2
*Psychiatric Service Facility Types in Study Sample*

| Service type | Number of facilities |
|---|---|
| Long-term inpatient locked units | 7 |
| Acute short-term inpatient locked units | 14 |
| Acute short-term inpatient unlocked units | 1 |
| Partial hospitalization programs | 5 |
| Crisis stabilization/respite programs | 6 |
| Crisis/emergency evaluation | 19 |
| Outpatient milieu programs (e.g., day treatment) | 16 |
| Outpatient therapy programs | 135 |
| Outpatient assessment and referral services | 5 |
| Community living/supported housing | 3 |
| Residential programs | 55 |
| Employee assistance programs | 2 |
| Case management | 2 |
| Other | 21 |

*Note*. Reported frequencies are for the total study sample.

influence the results and their generalizability. Given the size of the sample, tests were conducted and patterns examined for those variables/items with greater than 5% missing (Tabachnick & Fidell, 2007). These indicated that the data were missing at random with missing status having no meaningful effect on the results.[1]

Sample 1 was used to develop a baseline model. Parent-reported responses to the 103 items were correlated, and the resulting matrix was submitted to principal-components analysis (PCA) followed by oblique (Direct Oblimin) rotations. The optimal number of components to be retained was determined by the criterion of eigenvalues greater than 1, supplemented by the scree test and the criterion of interpretability (Cattell, 1966; Tabachnick & Fidell, 1996). Items that did not load on at least one component greater than 0.45, and components with fewer than two items were trimmed from the model. This relatively high cutoff (i.e., .45) was chosen deliberately to achieve the primary goal of scale reduction and to reduce the number of cross-loading items.

Sample 2 was then used to test the resulting model in a CFA framework and to revise the model using modification indices and fit diagnostics. Analyses were conducted using LISREL 8.80 (Jöreskog & Sörbom, 2006). Goodness of fit was

[1]We found that the demographic variable *income* and scale items related to school problems were above the 5% threshold. We performed a series of univariate tests for mean differences in factor scores between missing and nonmissing cases. Some of these tests were significant; however, this was likely because of the sample size. As suggested by Tabachnick and Fidell (2007), we also calculated effects sizes ($\eta^2$), which were extremely low (below .002), and this led us to conclude that these differences were not meaningful. Separate variance *t* tests of the school problem items indicated that these items were systematically missing in relation to one another. Importantly, when filling out the Child TOP, parents were instructed to skip the school items if their child was not presently attending school. As such, the test for missing completely at random was significant, yet predictable. Selecting out those cases for young children not yet in school (below age 5) resulted in a substantial percent decrease in missingness. According to Tabachnick and Fidell (2007), missing at random can still be inferred if the items are not missing completely at random but the missingness is predictable, which was the present case. As an added step, we once again compared missing and nonmissing cases in tests of mean differences. Once again, the effects sizes indicated that the differences were not meaningful. Based on this information, we concluded that items were not missing in way that would compromise subsequent analyses or the generalizability of the results.

evaluated using both absolute and relative indices, including the root mean square error of approximation (RMSEA), its 90% confidence interval (90% CI; cf. MacCallum, Browne, & Sugawara, 1996), and the RMSEA's associated probability value (Browne & Cudek, 1993), comparative fit index (CFI), and Tucker-Lewis index (TLI). Acceptable model fit was defined by the following criteria: RMSEA ($<0.08$, 90% CI, $<0.08$), CFI ($>0.90$), TLI ($>0.90$; Bentler, 1990; Hu & Bentler, 1999). Good model fit was defined by the following criteria RMSEA ($<0.05$), RMSEA associated $p$-value ($>.50$; Jöreskog & Sörbom, 1996), CFI ($>0.95$), and TLI ($>0.95$; Bentler, 1990; Hu & Bentler, 1999). Multiple indices were used because they provide different information about model fit (i.e., absolute fit, fit adjusting for model parsimony, fit relative to a null model); used together, these indices provide a more conservative and reliable test of the solution (Brown, 2006; Jaccard & Wan, 1996). Most of the revised models were nested; in these situations, comparative fit was evaluated by $\chi^2$ differences tests ($\chi^2_{\text{diff}}$) and the interpretability of the solution. The final model that resulted from sample 2 procedures was then replicated in a large, independent sample using CFA and employing the fit criteria stated above.

Finally, based on the pattern of correlations observed between the scales estimated as first-order factors in the CFA (see Table 5), a higher-order factor model was estimated.[2] In implementing this step, we returned to sample 2 to use PCA to explore the appropriate number of higher-order factors to estimate. The resulting higher-order model suggested by this PCA was then estimated in sample 3 within a CFA framework.

## Results

Based on the above criteria, 12 and 13 component models were extracted and reviewed from the PCA conducted on sample 1. Both orthogonal and oblique rotations were explored, with only minor differences found in component loadings. With the assumption that these components were related to each other, the oblique rotation was chosen. Items were trimmed because of insufficient loadings. In total, 45 of the 103 items were dropped and the final 58 items were again analyzed with oblique rotations. The final model appeared to be most parsimonious with 12 components accounting for 58% of the variance.

This model was then tested using structural equation modeling in an initial CFA (sample 2). Although this model met the conservative multiple-index fit criteria for acceptable fit (see Table 3), fit diagnostics indicated that the model could be improved to achieve the criteria for good fit. Through exploring all possible sources of strain (potential cross loadings, method effects, over-factoring or under-factoring, and minor factors), a series of steps were taken to improve the model, now using the CFA framework in an exploratory fashion. With each modification, the $\chi^2_{\text{diff}}$ was significant ($p<.001$). During this process the thirteen factor solution produced less strain. Essentially, this meant breaking up the questions that loaded on a violence/suicide scale into two separate factors. In addition, 10 more items were eliminated from the model because of relatively low-factor loadings and the item having more than one correlated error with items on other factors. In these cases, dropping the item from the model improved overall model fit. The model was also improved by freeing seven items to cross-load on other factors (see Table 4). The absolute value for the standardized regression weights for these cross-loaded items ranged from

[2]We thank an anonymous reviewer for the suggestion to perform these analyses.

Table 3
*CFA Validation and Cross-Validation*

| Model description | *n* | $\chi^2$ | df | TLI | CFI | RMSEA | RMSEA 90% CI upper | RMSEA *p* value |
|---|---|---|---|---|---|---|---|---|
| Initial model (12-factors) | 2449 | 11,898.673 | 1529 | .934 | .939 | .053 | .054 | 0.000 |
| Final model (13-factors) | 2449 | 4100.995 | 987 | .965 | .970 | .036 | .037 | 1.000 |
| Final model (confirmation) | 7267 | 9904.609 | 987 | .969 | .973 | .035 | .036 | 1.000 |
| Second-order model | 7267 | 15320.969 | 1055 | .956 | .958 | .043 | .044 | 1.000 |

*Note*. CFA, confirmatory factor analysis; df, degrees of freedom; TLI, Tucker-Lewis Index; CFI, comparative fit index; RMSEA, root mean square error of approximation; CI, confidence interval.

0.095 to 0.364. Finally, seven correlated errors were mapped into the model. Two were mapped because of item juxtaposition (46–47, 86–87), four were mapped because of item content similarity (10–86, 10–14, 14–86, 30–51), and one for both reasons (23–24). All modifications to the model were made based on both strain indices and the conceptual interpretation of the findings.

Sample 3 was used to replicate the final model developed with sample 2 and showed excellent and consistent model fit criteria across all indices. Taken together, there is strong support for the stability and strength of these factors. Results are summarized in Table 4 and demonstrate excellent model fit with no significant strains. The factor names and intercorrelations are listed in Table 5.

Given this pattern of intercorrleations among the final 13 first-order factors, we attempted to extend the analyses by estimating a model with second-order factors accounting for the relationship between these variables. Following the same procedures outlined above, we subjected the final 13 scales to an exploratory PCA with oblique rotation, which resulted in an interpretable three component solution.[3] Table 6 shows the loadings of the thirteen scales on these three components. We chose to label these three factors, internalizing, externalizing, and adjustment, based on the content of the scales whose principle loadings were on each factor. The correlation of internalizing with externalizing and adjustment was $r = 0.22$ and $-0.17$, respectively. The correlation between externalizing and adjustment was $r = -0.25$. Next, we attempted to fit this higher-order structure suggested by this PCA to the existing model in the CFA framework. The correlated standard errors were again constrained in estimating this model because of difficulties with convergence. The fit statistics for the second-order model can be found in Table 3. Although the model resulted in a significant $\chi^2$ difference test ($\chi^2_{\text{diff}}$[68], 5416.36; $p = <.001$), all other fit statistics remained within the range of good model fit.

[3]As the reader may note, we chose to use differing levels of factor loadings for constructing the first-order and second-order factor models (i.e., .45 vs. .30), which is because of differences in the goals at each phase of analysis. In an effort to reduce the total number of items from the initial 108 to a number that was manageable for daily use of the measure, we chose the more conservative .45 to achieve the primary goal of scale reduction and to reduce the number of cross-loading items. However, the goal of the second-order model was to estimate a small number of theoretically meaningful higher-order factors for augmented clinical and research utility without reducing the content of the measure; therefore, we used a more traditional .30.

Table 4
*Final CFA Standardized Regression Weights*

| | SEPAX | SEXWR | UNEAT | ADHDS | SLEEP | RSLNT | DEPRS | PSYCS | VIOLN | SUICD | CNDCT | ASRTV | ELMAT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Been afraid of being alone or did not want to be alone | 0.743 | | | | | | | | | | | | |
| Been very distressed when away from mother/caretaker | 0.554 | | | | | | | | | | | | |
| Needed someone (mother/caretaker) nearby in order to fall asleep | 0.600 | | | | | | | | | | | | |
| Made sexual comments | | 0.819 | | | | | | | | | | | |
| Caused you to worry about his/her sexual activity | | 0.710 | | | | | | | | | | | |
| Been a picky eater | | | 0.687 | | | | | | | | | | |
| Eaten too little | | | 0.597 | | | | | | | | | | |
| Eaten a variety of foods (vegetables, fruits, grains, meat) | | | −0.520 | 00.364 | | 0.335 | | | | | | | |
| Had trouble finishing things | | | | 0.652 | | | | | | | | | |
| Lost things | | | | 0.537 | | | | | | | | | |
| Had trouble staying still | | | | 0.691 | | | | | | | | | |
| Had trouble waiting | | | | 0.735 | | | | | | | | | |
| Had trouble paying attention in class | | | | 0.614 | | | | | | | | | |
| Been slow at completing homework | | | | 0.461 | | | | | | | | | |
| Had trouble getting back to sleep in the night | | | | | 0.797 | | | | | | | | |
| Woke up during the night (excluding trips to bathroom) | | | | | 0.789 | | | | | | | | |
| Had trouble falling asleep | | | | | 0.669 | | | | | | | | |
| Had nightmares | 0.267 | | | | 0.474 | | | | | | | | |
| Followed rules to your satisfaction | | | | | | 0.768 | | | | | | | |
| Done what he/she was asked to do | | | | | | 0.776 | | | | | | | |
| Been easy to live with | | | | | | 0.538 | −0.160 | | | | | | |
| Gotten along well with others | | | | | | 0.520 | | | | | | | |
| Done what was asked of him/her | | | | | | 0.662 | | | | | | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Been able to complete something after complaining that it was boring | | 0.449 | | | | | | | |
| Looked to share interests and exciting things with others | | 0.305 | | | | | | | |
| Been stuck in a certain mood and been unable to change | | | 0.708 | | | | | | |
| Had trouble looking people in the eye when talking to them | 0.292 | | 0.316 | | | | | | |
| Shown little emotion when you expected some type of reaction | | | 0.599 | | | | | | |
| Had little or no interest in things that were enjoyable before | | | 0.660 | | | | | | |
| Been able to talk but refused to do so | | | 0.476 | | | | | | |
| Not wanted to be touched | | | 0.518 | | | | | | |
| Looked down or depressed | | | 0.662 | | | | | | |
| Seen things that were not there | | | | 0.861 | | | | | |
| Heard things that were not there | | | | 0.864 | | | | | |
| Had desires to seriously hurt someone | | | | | 0.852 | | | | |
| Had thoughts of killing someone else | | | | | 0.795 | | | | |
| Seriously hurt someone | | | | | 0.608 | | | | |
| Physically hurt a person or an animal | 0.225 | | | | 0.542 | | | | |
| Hurt himself/herself | | | | 0.095 | | 0.645 | | | |
| Thought about killing himself/herself or wished to be dead | | | | | | 0.803 | | | |
| Stolen or shoplifted | | | | | | | 0.602 | | |
| Run away | | | | | | | 0.559 | | |
| Had trouble with the police | | | | | | | 0.652 | | |
| Seemed scared around people | | | | | | | | 0.671 | |
| Been too shy | −0.192 | | | | | | | 0.785 | |
| Had trouble standing up for himself/herself | | | | | | | | 0.585 | |
| Soiled underwear | | | | | | | | | 0.605 |
| Wet clothes or bed | | | | | | | | | 0.693 |

*Note*. CFA, confirmatory factor analysis; SLEEP, sleep functioning; SEPAX, separation anxiety; ADHDS, ADHD symptoms; CNDCT, conduct; VIOLN, violence; RSLNT, resiliency; ASRTV, assertiveness; UNEAT, under eating; DEPRS, depression; ELMAT, elimination problems; PSYCS, psychosis; SEXWR, sexual worry behavior; SUICD, suicide.

Table 5
*Factor Correlation Matrix*

| ABV | FULL NAME | SLEEP | SEPAX | ADHDS | CNDCT | VIOLN | RSLNT | ASRTV | UNEAT | DEPRS | ELMAT | PSYCS | SEXWR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SLEEP | Sleep functioning | | | | | | | | | | | | |
| SEPAX | Separation anxiety | .468 | | | | | | | | | | | |
| ADHDS | ADHD symptoms | .336 | .303 | | | | | | | | | | |
| CNDCT | Conduct | .132 | .060 | .206 | | | | | | | | | |
| VIOLN | Violence | .295 | .231 | .447 | .311 | | | | | | | | |
| RSLNT | Resiliency | .135 | .051 | .386 | .182 | .309 | | | | | | | |
| ASRTV | Assertiveness | .228 | .319 | .147 | .018 | .089 | −.027 | | | | | | |
| UNEAT | Under eating | .222 | .177 | .101 | .043 | .110 | .235 | .153 | | | | | |
| DEPRS | Depression | .392 | .324 | .478 | .276 | .443 | .344 | .303 | .237 | | | | |
| ELMAT | Elimination problems | .132 | .203 | .232 | .078 | .196 | .121 | .105 | .033 | .123 | | | |
| PSYCS | Psychosis | .332 | .276 | .248 | .221 | .315 | .095 | .177 | .103 | .333 | .126 | | |
| SEXWR | Worrisome sexual behavior | .156 | .106 | .254 | .288 | .311 | .169 | .006 | .056 | .264 | .107 | .196 | |
| SUICD | Suicide | .312 | .218 | .271 | .256 | .475 | .195 | .166 | .147 | .477 | .083 | .385 | .266 |

*Note.* ABV, scale abbreviation. SEPAX, separation anxiety; ADHDS, attention deficit hyperactivity symptoms; CNDCT, conduct; VIOLN, violence; RSLNT, resilience; ASRTV, assertiveness; UNEAT, under eating; DPRS, depression; ELMAT, elimination problems; PSYCS, psychosis; SEXWR, worrisome sexual behavior; SUICD, suicide.

Journal of Clinical Psychology DOI: 10.1002/jclp

Table 6
*Principle Component Analysis of Child TOP First-Order Scales*

| Scale | Internalizing | Externalizing | Adjustment |
|---|---|---|---|
| Assertiveness | **.713** | −.111 | −.006 |
| Separation anxiety | **.688** | −.030 | −.122 |
| Under eating | **.530** | .037 | .078 |
| Sleep functioning | **.500** | .293 | −.124 |
| Depression | **.389** | **.436** | −.262 |
| Conduct | −.195 | **.695** | .039 |
| Suicide | .124 | **.679** | −.022 |
| Worrisome sex behavior | −.030 | **.606** | −.065 |
| Psychosis | .296 | **.525** | .170 |
| Violence | .000 | **.519** | **.398** |
| Resiliency | .276 | −.163 | **.763** |
| ADHD symptoms | .239 | .185 | **−.603** |
| Elimination problems | .129 | −.177 | **−.588** |

*Note*. TOP, Treatment Outcome Package; ADHD, attention deficit hyperactivity disorder; $n = 2{,}449$; loadings of over .300 are in bold.

## Discussion

The child version of the TOP was developed to assess a broad range of behavioral health functional and symptom domains. Designed for clinical and research use in naturalistic settings, it addresses a large number of recommendations that have been made by experts in youth assessment and treatment (Achenbach, 2005; Kazdin, 2005; Mash & Hunsley, 2005), as well as in psychotherapy in general (Horowitz et al., 1997). In the present study, the 103 items constituting a previous version of the Child TOP were submitted to principle components and CFAs based on large samples of diverse patients. After a principle components and initial CFA (which led to the removal of 45 and 10 items, respectively), a final CFA conducted with the retained 48 items demonstrated excellent fit to a 13-factor scale model.

A number of these 13 factors share questions and patterns similar to the factors identified in the adult version of the TOP (Kraus et al., 2005). These factors include, sleep functioning, violence, psychosis, and suicide. The factor for depression contained items similar to the adult construct, such as "Had little or no interest in things that were enjoyable before" and "Looked down or depressed." Other questions in the child depression factor, however, included questions like "not wanted to be touched" and "shown little emotion when you expected some type of reaction." These questions, which are likely to be especially relevant to child development, are consistent with an attachment theory of depression (Bowlby, 1980).

A number of parallels between the development of the Adult and Child TOP questionnaires are important to note. First, although PCAs of both scales suggested combining suicidality and violence into a single factor, CFAs produced less strain with them split, and improved still further by allowing one important item to cross load. Along with the relatively high correlation between suicide and violence factors (.475), these findings point to important clinical and research implications when evaluating self-injury, violence, and suicide for all age groups.

Second, the violence scales on both adult and child versions were improved by eliminating mild temper symptoms like angry outbursts and tantrums and revising the scale to become a more serious violence scale focusing on inflicting pain on

others and planning homicide. The clinical and etiological implications of this should be more fully explored.

A number of factors included symptoms specifically related to DSM-IV childhood disorders, such as separation anxiety and attention deficit hyperactivity disorder symptoms. We found it interesting that other factors indicate that the Child TOP assesses difficulties that are not directly or specifically captured by discrete diagnostic categories, e.g., assertiveness, worrisome sexual behavior. In addition, the child version (as well as the adult version, see Kraus et al., 2005) is not restricted to manifestations of distress or impairment. Specifically, the resiliency factor appears to capture a separate and unique factor of childhood strengths, which is of particular note given modern concepts of resilience and adaptability (Luthar, 2003). Thus, by assessing both DSM and non-DSM types of distress, as well as markers of psychological resources or health, the Child TOP addresses numerous dimensions that are relevant to comprehensive case formulations and treatment plans. Finally, second-order principle components analysis suggested a solution with three second-order factors; however, as mentioned above, it resulted in a significant increase in the chi-square change in CFA models in these very large samples. Nevertheless, the alternative fit statistics suggest good fit for this higher-order CFA model. Furthermore, these factors are easily interpretable as being comprised of externalizing symptoms, internalizing symptoms, and adjustment behaviors. Although the first two of these factors reflect major dimensions of child psychopathology (Tackett & Krueger, 2005), the items loading on the adjustment factor can potentially be turned into a measure of adaptive functioning, which would be consistent with recommendations made by several experts in child psychopathology and treatment (Achenbach, 2005; Kazdin, 2005; Mash & Hunsley, 2005). We suspect that this higher-order model may vary across child age ranges (e.g., quite young to preteenage). For example, across all possible age groups in the sample, the Depression scale loaded on both the higher-order factors of internalizing and externalizing. We speculate that the Depression scale may load on one or the other depending on further differentiations of child age groups. Future analyses are planned that will examine the higher-order factor structure by age group.

Future analyses should be pursued to examine the second-order factor structure of the child version of the TOP, using more homogeneous groups than in the current study, as well as composite variables (i.e., scale scores). If validated, such a structure might lead some clinicians and researchers to use a shorter version of the instrument, for pragmatic (e.g., to save time for clients), methodological (e.g., to reduce number of tests when conducting small sample studies), and/or for conceptual purposes (to focus on clusters or categories of disorder such as externalizing and internalizing psychopathology). However, in line with the recommendations of the APA/SPR core battery conference (Horowitz et al., 1997) and several leading researchers in the field of youth psychopathology (e.g., Achenbach, 2005; Kazdin, 2005), we believe that one of the main advantages of the current version of the TOP is that it provides an assessment of a wide array of difficulties and strengths which, needless to say, can be very helpful to a comprehensive case formulation and appropriate treatment plan.

This study has both strengths and weaknesses. A limitation of the present study may be the oversimplification of the original Child TOP, eliminating items that provide a wealth of clinical information and feedback to therapists. However, when constructing a measure such as this to be used in naturalistic settings, a balance must be struck between clinical usefulness and psychometric properties and efficiency. The present study attempted to meet the goal of creating a scale short enough for

everyday use. This study also has several strengths. For example, this study was conducted with a large, representative clinical sample that was large enough to be split for cross-validation purposes. The use of both exploratory and confirmatory methods also helps to strengthen the reliability and validity of the findings. This study also examined a potential higher-order factor structure.

Future studies on the child version of the TOP should also include investigations of convergent, discriminant, concurrent, and predictive validity, analyses of measurement invariance in different age groups and genders, temporal stability and sensitivity to change, and its ability to differentiate between psychiatric clients and children in the general population. Of particular interest with respect to its sensitivity to change, future research on the large TOP outcome database should assess whether interventions that appear to improve the resiliency factor also facilitate positive outcomes on other symptom based factors. Finally, preliminary research on the Adult TOP (Stelk & Berger, 2009) suggests that the TOP has predictive validity in identifying patients who will be hospitalized within the next 6 months, a focus of investigation that should be extended to Child TOP data.

## References

Achenbach, T.M. (2005). Advancing assessment of children and adolescents: Commentary on evidenced-based assessment of child and adolescent disorders. Journal of Clinical Child and Adolescent Psychology, 34, 541–547.

American Psychiatric Association (1994). Diagnostic and statistical manual of mental disorders (4th ed.). Washington, DC: American Psychiatric Association.

Bentler, P.M. (1990). Comparative fit indexes in structural models. Psychological Bulletin, 107, 238–246.

Bowlby, J. (1980). Attachment and loss: Vol. III. Loss: Sadness and depression. New York: Basic Books.

Brown, T.A. (2006). Confirmatory factor analysis for applied research. New York: Guilford.

Browne, M.W., & Cudek, R. (1993). Alternate ways of assessing model fit. In K.A. Bollen & J.S. Long (Eds.), Testing structural equation models (pp. 136–162). Newbury Park, CA: Sage.

Burisch, M. (1986). Methods of personality inventory development—A comparative analysis. In A. Angleitner & J.S. Wiggins (Eds.), Personality assessment via questionnaires (pp. 109–120). New York: Springer-Verlag.

Cattell, R.B. (1966). The scree test for the number of factors. Multivariate Behavioral Research, 1, 245–276.

Floyd, F.J., & Widaman, K.F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. Psychological Assessment, 7, 286–299.

Hermann, R. (2003). Risk adjustment for mental health care. In L. Iezzoni (Ed.), Risk adjustment for measuring health care outcomes (3rd ed.). Ann Arbor, MI: Health Administration Press.

Horowitz, L.M., Lambert, M.J., & Strupp, H.H. (Eds.). (1997). Measuring patient change in mood, anxiety, and personality disorders: Toward a core battery. Washington, DC: American Psychological Association Press.

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

Jaccard, J., & Wan, C.K. (1996). LISREL approaches to interaction effects in multiple regression. Thousand Oaks, CA: Sage.

Jöreskog, K.G., & Sörbom, D. (1996). LISREL 8: User's reference guide. Chicago, IL: Scientific Software International.

Jöreskog, K.G. & Sörbom, D. (2006). LISREL 8.8 for Windows [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

Kazdin, A.E. (2005). Evidence-based assessment for children and adolescent: Issues in measurement development and clinical application. Journal of Clinical Child and Adolescent Psychology, 34, 548–558.

Kraus, D.R., Seligman, D., & Jordan, J.R., (2005). Validation of a behavioral health treatment outcome and assessment tool designed for naturalistic settings: The treatment outcome package. Journal of Clinical Psychology, 61, 285–314.

Luthar, S.S. (Ed.). (2003). Resilience and vulnerability. Adaptation in the context of childhood adversities. New York: Cambridge University Press.

MacCallum, R.C., Browne, M.W., & Sugawara, H.M. (1996). Power analysis and determination of sample size for 575 covariance structure modeling. Psychological Methods, 1, 130–149.

Mash, E.J., & Hunsley, J. (2005). Special section: Developing guidelines for the evidence-based assessment of child and adolescent disorders. Journal of Clinical Child and Adolescent Psychology, 34, 362–379.

Stelk, W., & Berger, M. (2009, June). Predictive modeling: Using TOP clinical domain items to identify adult medicaid recipients at risk for high utilization of behavioral health services in a managed care provider network. Paper presented at the Society for Psychotherapy Research conference, Santiago, Chile.

Tabachnick, B.G., & Fidell, L.S. (1996). Using multivariate statistics (3rd ed.). New York: Harper Collins College Publishers.

Tabachnick, B.G., & Fidell, L.S. (2007). Using multivariate statistics (5th ed.). New York: Allyn and Bacon.

Tackett, J.L., & Krueger, R.F. (2005). Interpreting personality as a vulnerability to psychopathology: A developmental approach to personality-psychopathology relationship. In B.L. Hankin & J.R.Z. Abela (Eds.), Development of psychopathology: A vulnerable-stress perspective (pp. 199–214). Thousand Oaks, CA: Sage Publications.

Wiggins, J.S. (1973). Personality and prediction: Principles of personality assessment. Menlo Park, CA: Addison-Wesley.