# Development of a Change-Sensitive Outcome Measure for Children Receiving Counseling

Scott T. Meier, James L. McDougal and Achilles Bardos

*Canadian Journal of School Psychology* 2008; 23; 148 originally published online Mar 7, 2008;
DOI: 10.1177/0829573507307693

**Canadian Journal of School Psychology**
Volume 23 Number 2
December 2008 148-160
© 2008 Sage Publications
10.1177/0829573507307693
http://cjsp.sagepub.com
hosted at
http://online.sagepub.com

# Development of a Change-Sensitive Outcome Measure for Children Receiving Counseling

Scott T. Meier
*University at Buffalo*
James L. McDougal
*State University of New York at Oswego*
Achilles Bardos
*University of Northern Colorado*

***Abstract:*** Contemporary testing standards place test purpose as the central focus during test development and subsequent use. This study describes the development of a measure for children designed explicitly to measure change resulting from psychosocial interventions. Parents completed the outcome measure for 896 elementary school-age children receiving psychotherapy interventions from community mental health agencies. Scales formed with change-sensitive items evidenced adequate reliability estimates and larger effect sizes than scales composed of the original item pool. When asked by stakeholders such as parents, principals, and school boards to investigate the effectiveness of provided interventions, school psychologists should consider the use of change-sensitive measures that are sensitive to small and moderately sized treatment effects.

***Résumé:*** De nos jours, il est pratique courante d'accorder, au cours de l'élaboration du test psychologique et de son utilisation ultérieure, une place centrale à l'objectif qu'il vise. La présente étude aborde l'élaboration d'un test pour enfants ayant pour objectif de mesurer d'une manière explicite les changements occasionnés par les interventions psychosociales. Des parents ont participé à ce test qui concernait 896 enfants d'âge primaire bénéficiant de l'aide de psychothérapeutes d'organismes communautaires de santé mentale. Les tests contenant des items aptes à détecter les changements attribuables au counseling ont prouvé leur efficacité et permettent de mieux mesurer ces changements que les tests constitués des items initiaux. Lorsque les intervenants, par exemple les parents, les chefs d'école et les conseils scolaires, leur demandent de déterminer l'efficacité des interventions offertes, les psychologues scolaires devraient envisager de recourir aux tests contenant des items aptes à détecter les changements attribuables au counseling.

**Authors' Note:** Please address correspondence to Scott Meier, Counseling, School, and Educational Psychology, 409 Baldy Hall, SUNY Buffalo, Buffalo, NY 14209; e-mail: stmeier@buffalo.edu.

**Keywords:** *testing; measurement; outcome assessment; counseling; psychotherapy; children and adolescents*

Psychologists interested in assessing children's and adolescents' changes resulting from psychosocial interventions face a significant obstacle: Many of the most popular scales, such as the Child Behavior Checklist and Conners's Rating Scales (Achenbach, 1994; Conners, 1994), are lengthy instruments designed primarily for diagnostic and screening purposes. As Hill and Lambert (2004) noted, "most outcome measures have not been developed with an eye toward choosing items that are sensitive to change, and little is known about this aspect of test validity" (p. 117). Contemporary standards place testing purpose as the central focus during test development; testing purpose drives the test specification process, which includes decisions about the desired psychometric properties of items (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Because the primary purpose of an outcome measure is to assess client changes in response to an intervention, an instrument's sensitivity to change is directly related to its construct validity (Vermeersch et al., 2004).

Research on outcome measures' sensitivity to change has demonstrated that such tests differ considerably in their ability to detect the types and amount of change following therapy. Lambert, Hatch, Kingston, and Edwards (1986) compared the Zung, Beck, and Hamilton depression scales as outcome measures of mental health interventions and found "that rating devices can by themselves produce differences larger than those ordinarily attributed to treatments" (p. 58). One explanation for differences in the sensitivity of measures to detect intervention effects is that test items are differentially sensitive (Meier, 1997). That is, individual items may be more or less able to detect the multiple effects of any given program or intervention.

In this vein, Meier (1997, 1998, 2000) proposed a set of intervention item selection rules (IISRs) designed to identify intervention-sensitive items during either test construction or subsequent item evaluations. This approach assumes that (a) items differ along a trait–state continuum and (b) different test construction and item analysis procedures are necessary to select items and create scales that reflect intervention effects. Table 1 contains a brief description of the IISRs, and a more detailed description of the methods used to assess each guideline is provided in the Results section. The fourth IISR is the pivotal guideline, specifying that baseline assessments at intake should be compared with one or more assessment periods following an intervention (Cronbach et al., 1980; Speer & Newman, 1996). Initial research has found differences between the psychometric properties of scales developed with intervention-sensitive guidelines and traditional item selection guidelines such that the former demonstrate larger effect sizes (ESs) after counseling interventions (Meier, 1998, 2000, 2004; Vermeersch, Lambert, & Burlingame, 2000; Vermeersch et al., 2004; Weinstock & Meier, 2003).

**Table 1**
**Brief Description of Intervention Item Selection Rules**

| Rule | Description |
|---|---|
| 1 | Ground scale items in theoretical and empirical literature relevant to applicable interventions and target problems |
| 2 | Aggregate at appropriate levels |
| 3 | Assess range of item scores at pretest |
| 4 | Detect change in an item's score after an intervention |
| 5 | Assess whether change occurs in the expected direction |
| 6 | Examine whether differences in change exist between intervention and comparison groups |
| 7 | Examine whether intake differences exist between comparison groups |
| 8 | Examine relations between item scores and systematic error sources |
| 9 | Aggregate selected items into scale(s) and cross-validate |

This study uses IISRs to develop a new scale, the Parent Elementary form of the Behavioral Intervention Monitoring Assessment System (PE-BIMAS; Meier, McDougal, & Bardos, in press). The PE-BIMAS's primary purpose is to assess change in children receiving psychosocial interventions in field settings, a combination rarely studied in psychotherapy research (Kazdin, 2000). The purpose of the current study is to (a) evaluate PE-BIMAS items according to IISR guidelines, in preparation for creating aggregate scales, and then (b) to evaluate the psychometric properties, particularly the change sensitivity, of the resulting scales.

# Method

## Procedure

All parents or guardians of children and adolescents who sought counseling at two multiple-branch community mental health agencies were asked to complete an intake version of the PE-BIMAS at their first visit. Parents completed an informed consent form that indicated that the purpose of the study was program evaluation of the counseling services. The agencies provide individual, family, and group interventions to children, adolescents, and families in urban, suburban, and rural communities. These agencies work closely with school personnel, including school psychologists, and participated in this research in exchange for program evaluation reports required by school administrators and external agencies who provided funding support. Because these agencies had experienced difficulty obtaining outcome assessments at termination, they adopted a procedure where follow-up forms were administered at quarterly intervals to all clients receiving services (i.e., approximately four times per year; cf. Lewis & Magoon, 1987). Thus, a unique intake form was collected per client, but multiple follow-up forms were possible; the mean amount of time between intake and follow-up in the PE-BIMAS database was 7.2 months (*SD* = 5.1, range 1 to 38 months).

A parent or another adult familiar with the child first provided brief demographic information about the child. Raters were then instructed to report how often each of the behaviors, events, or situations described in PE-BIMAS items occurred during the past week. Parents answered each item by checking a box on the following scale: *never* (0 instances in the past week, coded as 0), *rarely* (1 instance, coded as 1), *sometimes* (2-3 instances, coded as 2), and *often* (3-7 instances, coded as 3).

## Participants

During a period of 5 years, this procedure resulted in the compilation of a database of 2,002 intake and 2,588 follow-up forms. Of that total, 896 distinct clients (45% of intakes) had completed at least one intake and one follow-up form; 1,101 (55% of intakes) completed an intake but no follow-up, and 250 (12% of intakes) completed two or more follow-ups. Of those 896 clients with complete data, 742 (83%) reported data about client gender (66% identified as male, 34% as female), and 706 (79%) reported client race (60% Caucasian, 20% African American, 8% Hispanic, 4% Biracial, 2% Native American, 3% Asian, and 3% Other). A total of 601 parents (67%) reported school grade (12% in Kindergarten, 14% in Grade 1, 13% in Grade 2, 14% in Grade 3, 19% in Grade 4, 15% in Grade 5, 11% in Grade 6, and the remaining 2% in Grades 7-10), and 612 (68%) reported their relationship to client (67% as mother, 7% as father, and 26% as other). Students' presenting issues included peer difficulties, family conflict, illegal activities that brought children into court, emotional and physical problems, lack of self-confidence, child abuse, and hyperactivity.

## Measures

Intended to function as a comprehensive measure of outcome, the PE-BIMAS contains items assessing broad content domains pertinent to children's functioning and outcome information. The items were developed on the basis of a review of the literature on children's behavior problems (Stiffman, Orme, Evans, Feldman, & Keeney, 1984), other measures of children's distress and functioning (Meier, 1998), and suggestions from mental health professionals. The review identified categories including hiding thoughts from others, deviancy, internalizing behaviors (e.g., anxiety, depression), externalizing behaviors (e.g., fighting), and problems in cognitive, social, and academic functioning (Stiffman et al., 1984). This process resulted in the creation of a 32-item scale (labeled the Original scale). To avoid acquiescence and criticalness biases, items were worded in both positive (labeled the Strengths scale) and negative (labeled the Distress/Problems scale) directions.

Because of staff reports of parents' reluctance to complete what they perceived as lengthy questionnaires, agency administrators asked that the scale be reduced to a one-page form that included basic demographic information. The 32-item Original scale was thus reduced to a 16-item scale on the basis of item analyses that evaluated item–total correlations, change sensitivity, and content validity. After several periods of data collection, staff requested that other items, principally related to family functioning, be added to the scale and several others dropped. Thus, three versions of the PE-BIMAS were eventually used, with 16, 17, and 17 total items; 14 items overlapped these versions. For the PE-BIMAS database, these changes resulted in data collection for 19 total items, with 5 items (sleepy or tired, starts conversations, well behaved at home, family members fight, and limits set with children) possessing a smaller sample size than the overlapping 14 items.

## Results

As with most test construction procedures, the intended products of the IISR analyses are multi-item scales. The results of each IISR step analysis are used when deciding which items to include in change-sensitive scales. Items are not deleted or added to scales during any particular step, but results from each step are considered when creating scales. Information regarding each IISR, as applied to the PE-BIMAS database, is described below. At IISR 9, item-level results from each previous IISR were used to create multi-item scales. To address possible chance effects, scales were cross-validated; to do so, study participants were randomly assigned to one of two groups, resulting in Subsample A and Subsample B, both with a sample size of 448. Randomization succeeded in creating two roughly equivalent groups, as shown for the demographic variables of gender (Subsample A had 69% males, Subsample B 64%), relation to client (66% mother in A, 68% mother in B), and race (62% Caucasian and 20% African American in A compared to 58% and 19%, respectively, in B). As described below, the total sample or both subsamples were also used in each IISR analysis, depending on the rationale for that guideline as well as the size of the available sample.

1. *Ground items in previous research and theory.* Relevant research and theory provide a context for understanding the meaning of changing scores on an intervention-sensitive measure. In the area of child and adolescent psychotherapy, Kazdin (2000) noted that more than 1,000 controlled studies of psychosocial interventions for children and adolescents exist. Kazdin maintained that because ESs for all interventions averaged about .70 for children and adolescents, maturation alone cannot account for such gains. Meta-analytic studies also indicate that adolescents appear to benefit more from psychotherapy than children, although most of the difference can be attributed to the benefits received by adolescent girls (Weisz, Huey, & Weersing, 1998). Applied to this study, these findings suggest that (a) some PE-BIMAS items

should evidence positive change but that these effects will be small and (b) girls may be more likely to show positive change than boys.

2. *Aggregate items at an appropriate level.* Because an item response contributed by an individual on one occasion may be influenced by random error (Messick, 1989), item responses should first be aggregated across individuals before further analyses are conducted. Similarly, test developers have long recognized that aggregation of individual item responses into scales increases the reliability and validity of measurement of the studied construct. Intervention-sensitive items are not aggregated across occasions, however, but summed across individuals and items. As was done in this study, item scores are then compared across time periods in which interventions take place to determine if change effects are present at the level of aggregated item responses.

3. *Assess range of item scores at pretest.* Ceiling and floor effects inhibit detection of desired changes in intervention-sensitive tests because they can restrict the potential range of scores. In this study, a ceiling effect occurred when an item's standard deviation was added to the item mean and the resulting sum exceeded the highest value of the scale (3); a floor effect occurred when the item's standard deviation was subtracted from the item mean and the result was less than the bottom range of the scale (0). Three Strengths items in both subsamples had a ceiling effect: communicates clearly, starts conversations, and limits set with children. No floor effects were found.

4. *Items should evidence change in intervention conditions.* Intervention-sensitive items should demonstrate change, from baseline to follow-up periods, with clients who receive psychosocial interventions (cf. Cronbach et al., 1980). For the current study, paired *t* tests were computed to examine change in item scores from intake to follow-up. Because these analyses are exploratory in nature, and the expected effects at the level of an individual item are likely to be small, an $\alpha$ level of .10 was set to detect statistically significant change (cf. Meier, 2000). As shown in Table 2, 12 of 19 items evidenced statistically significant change in one or both subsamples: controls temper, pays attention to speakers, stays out of trouble, communicates clearly, shares thinking, feels depressed, behaves differently, acts impulsively, fights with others, family members fight, lies or cheats, and gets failing grades.

5. *Items should evidence change in the theoretically expected direction.* All 12 items that evidenced significant change in Table 2 improved from intake to follow-up. Although clients worsened in at least one subsample on the items makes friends easily, limits set with children, and helps with household tasks, these changes did not reach statistical significance.

6. *Evaluate item change in intervention and comparison groups.* Item change in intervention groups can be compared to change in items completed by available comparison groups. As noted above, Kazdin's (2000; see also Weisz, Weiss, Han, Granger, & Morton, 1995; Webster-Stratton, 1996) review found that girls evidence more improvement than boys as a result of psychosocial interventions. Meta-analytic

**Table 2**
**Results of Change Analyses for Parent Elementary–Behavioral Intervention Monitoring Assessment System Items**

| | Subsample A | | | Subsample B | | |
|---|---|---|---|---|---|---|
| Item Content by Scale | *M Diff* | *SE* | *t* | *M Diff* | *SE* | *t* |
| Strengths | | | | | | |
| Communicates clearly[b] | .13 | .04 | 3.47** | .05 | .03 | 1.33 |
| Stays out of trouble[b] | .13 | .04 | 3.08** | .05 | .04 | 1.13 |
| Controls temper[a] | .13 | .05 | 2.80** | .14 | .05 | 2.93** |
| Pays attention to speakers[a] | .12 | .04 | 2.78** | .16 | .04 | 4.07** |
| Well behaved at home | .07 | .05 | 1.51 | .05 | .05 | 1.02 |
| Shares thinking[b] | .06 | .04 | 1.36 | .09 | .04 | 2.10* |
| Limits set with children | .05 | .04 | 1.12 | −.03 | .04 | -0.63 |
| Starts conversations | .06 | .08 | 0.78 | .06 | .07 | 0.75 |
| Makes friends easily | .00 | .04 | 0.03 | −.06 | .04 | −1.51 |
| Helps with household tasks | −.00 | .05 | −0.02 | .02 | .04 | 0.45 |
| Distress/Problems | | | | | | |
| Feels depressed[a] | .22 | .05 | 4.69** | .14 | .05 | 3.17** |
| Behaves differently[a] | .13 | .04 | 2.91** | .18 | .04 | 4.01** |
| Family members fight[a] | .16 | .06 | 2.79** | .11 | .06 | 1.96* |
| Acts impulsively[a] | .10 | .04 | 2.59* | .14 | .04 | 3.38** |
| Fights with others[a] | .10 | .04 | 2.46* | .14 | .04 | 3.37** |
| Lies or cheats[a] | .12 | .05 | 2.43* | .11 | .04 | 2.55* |
| Gets failing grades[a] | .10 | .05 | 1.98* | .10 | .05 | 1.98* |
| Fidgets | .08 | .05 | 1.60 | .07 | .04 | 1.60 |
| Sleepy or tired | .01 | .09 | 0.10 | .07 | .08 | 0.83 |

Note: Items are arranged by descending absolute *t* value of Subsample A. A negative mean change score indicates that scores for this item worsened from initial to follow-up period.

a. Items evidence statistically significant change in both Subsample A and B.

b. Items evidence statistically significant change in either Subsample A or B.

$^{*}p < .10$. $^{**}p < .05$.

results, however, indicate that this finding may primarily be due to gains by adolescent girls (Weisz et al., 1998).

Because of missing reports of demographic information, gender analyses were not conducted by random subsample but for the entire sample of clients' parents who had provided information about gender ($n$ = 493 for boys and 249 for girls). Both boys and girls evidenced statistically significant improvement on six items: pays attention to speakers, controls temper, communicates clearly, feels depressed, acts impulsively, and fights with others. Boys evidenced more improvement than girls on four Distress/Problems items, changing on behaves differently, gets failing grades, lies or cheats, and family members fight. Girls demonstrated more improvement on three Strengths items, improving at a statistically significant level on stays out of trouble, shares thinking, and limits set with children. Boys also evidenced a statistically significant decrease on the Strengths item makes friends easily. Although differences in sample size likely contributed, these results indicate that boys and girls experienced some differential effects as a result of the psychotherapeutic interventions they received.

7. *Examine the equivalence of items scores at intake between groups.* In the PE-BIMAS data set, intake equivalence could be examined between the two randomly created subsamples A and B. Paired $t$ tests were used to assess differences between item means, and two items differed at intake: makes friends easily ($t = -1.78$, $p <$ .10) and family members fight ($t = -2.37$, $p < .05$). Overall, random assignment resulted in statistically equivalent groups, providing confidence that the subsequent cross-validation analyses (IISR 9) of intervention-sensitive items can be interpreted appropriately.

8. *Examine the relationship between scale items and systematic error sources.* No data were available for addressing this IISR.

9. *Aggregate selected items into scale(s) and cross-validate.* The preceding IISR analyses provide a basis for understanding the relevant properties of scale items and lay the foundation for subsequent decisions about inclusion in multi-item scales. Resulting scales are described below.

Because the PE-BIMAS scale was intended to reflect a broad range of potential problems and strengths, it could also be used as a screening instrument. This scale will be labeled the Total scale, consisting of the 14 overlapping items across the three versions used during test construction. The performance of all 19 items will be examined through two subscales: the 10-item Strengths and the 9-item Distress/Problems scales. Because not all of these items were present in all three versions of the PE-BIMAS, the sample size was reduced when calculating scale properties (see Table 3).

The items that evidenced no ceiling or floor effects (IISR 3), changed from intake to follow-up in the total sample (IISR 4), changed in the expected direction (IISR 5), did not show positive change in the control condition (IISR 6), showed equivalence between subsamples at intake (IISR 7), and were cross-validated (IISR 9) were combined into an eight-item Positive Change scale. Given the possibility of chance findings in these analyses, cross-validation of item selection results is important. As shown in Table 2, nine items have $t$ values with statistical significance below .10 for intake–follow-up differences for both randomly constructed subsamples: controls temper, pays attention to speakers, feels depressed, behaves differently, acts impulsively, fights with others, family members fight, lies or cheats, and gets failing grades. Consequently, these items can be considered cross-validated in terms of their sensitivity to change across subsamples.

Emphasizing the centrality of IISR 4, all 12 items that changed from intake to follow-up in either subsample will be grouped into a 12-item Positive Change scale. Containing a larger number of items, this broad-based scale has increased reliability

**Table 3**
**Descriptive Statistics, Coefficient αs, and Effect Sizes for Parent Elementary–Behavioral Intervention Monitoring Assessment System Scales**

| | Initial | | | | Follow-Up | | | Pooled | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Scale (No. Items) | *n* | *M* | *SD* | α | *n* | *M* | *SD* | α | *SD* | ES |
| Original scale (32) | 111 | 53.02 | 13.10 | .88 | 122 | 54.90 | 12.02 | .88 | 12.60 | .15 |
| Screening scales | | | | | | | | | | |
| Total (14) | 891 | 20.79 | 6.53 | .80 | 894 | 19.36 | 6.91 | .84 | 6.76 | .21 |
| Strengths (10) | 579 | 9.40 | 4.23 | .72 | 644 | 8.78 | 4.34 | .76 | 4.30 | .14 |
| Distress/Problems (9) | 581 | 14.65 | 4.37 | .72 | 647 | 13.62 | 4.59 | .76 | 4.51 | .23 |
| Intervention-sensitive scales | | | | | | | | | | |
| Positive Change (8) | | | | | | | | | | |
| Total sample | 892 | 13.46 | 4.26 | .73 | 895 | 12.39 | 4.47 | .78 | 4.40 | .24 |
| Boys' sample | 477 | 13.88 | 4.32 | .75 | 478 | 12.51 | 4.58 | .79 | 4.50 | .30 |
| Caucasian | 275 | 13.41 | 4.31 | .76 | 276 | 11.89 | 4.60 | .80 | 4.52 | .34 |
| Girls' sample | 238 | 12.89 | 4.21 | .71 | 240 | 11.98 | 4.53 | .77 | 4.39 | .21 |
| Positive Change (12) | | | | | | | | | | |
| Total sample | 579 | 18.20 | 5.67 | .77 | 646 | 16.70 | 6.18 | .82 | 5.99 | .25 |
| Boys' sample | 283 | 18.85 | 5.61 | .78 | 324 | 17.01 | 6.36 | .82 | 6.09 | .30 |
| Caucasian | 155 | 18.09 | 5.47 | .78 | 177 | 15.71 | 6.37 | .83 | 6.08 | .39 |
| Girls' sample | 139 | 17.69 | 5.90 | .76 | 159 | 15.98 | 6.48 | .83 | 6.27 | .27 |
| Boys' Change (10) | | | | | | | | | | |
| Total sample | 580 | 15.98 | 4.92 | .74 | 647 | 14.65 | 5.27 | .78 | 5.14 | .26 |
| Boys' sample | 284 | 16.57 | 4.82 | .74 | 324 | 14.84 | 5.42 | .79 | 5.22 | .33 |
| Caucasian | 155 | 16.09 | 4.69 | .74 | 177 | 13.82 | 5.49 | .80 | 5.25 | .43 |
| Girls' sample | 139 | 15.47 | 5.11 | .74 | 159 | 14.16 | 5.45 | .80 | 5.32 | .25 |
| Girls' Change (9) | | | | | | | | | | |
| Total sample | 579 | 11.74 | 4.02 | .68 | 644 | 10.75 | 4.25 | .73 | 4.17 | .24 |
| Girls' sample | 139 | 11.53 | 4.33 | .68 | 159 | 10.14 | 4.49 | .75 | 4.46 | .31 |
| Caucasian | 72 | 10.81 | 4.15 | .68 | 84 | 9.63 | 4.65 | .78 | 4.45 | .27 |
| Boys' sample | 283 | 12.20 | 3.97 | .68 | 322 | 11.01 | 4.42 | .75 | 4.25 | .28 |

Note: Statistics for the Screening scales were calculated with all available forms for the total sample, and statistics for the intervention-sensitive scales were calculated with the sample of clients whose parents provided both initial and follow-up ratings. Higher scores indicate greater frequency for all scales. A positive effect size (ES) indicates improvement from intake to follow-up.

compared to the 8-item change scale (see Table 3). And as described in IISR 6, the 10 items that displayed positive change with boys will be labeled the Boys' Change scale, and the 9 items that evidenced change with girls will be the Girls' Change scale.

As shown in Table 3, psychometric analyses at the scale level provide relevant reliability and validity information. Reliability estimates are high to average for the 32-item Original scale and the three Screening scales; the Screening scales αs had an intake mean of .75 and a follow-up mean of .79. Reliability estimates for the group of intervention-sensitive scales are similar: At intake, IISR scales' αs averaged

.75 and at follow-up, .80. Reliability values indicate that all scales have sufficient internal consistency for subsequent use and interpretation.

Because participant heterogeneity can decrease the power to detect treatment effects (Lipsey, 1998), scale results were computed and reported for the largest homogeneous subgroups in this data set (i.e., for total sample, boys and girls, and Caucasian boys). ES was calculated by subtracting the follow-up mean from the initial mean and then dividing that difference by the pooled (intake and follow-up) standard deviation. As shown in Table 3, the Original and Screening scales have lower ESs than intervention-sensitive scales. The mean ES for the three Screening scales equaled .19 (*SD* = .05), whereas the mean ES for the IISR scales across all subgroups was .29 (*SD* = .06). More homogeneous subgroups had higher ESs: Boys' IISR scales ranged from .30 to .33 and Caucasian boys' scales from .34 to .43. The 10-item Strengths scale (which includes 3 items with ceiling effects) and the 32-item Original scale (whose statistics are based on a smaller sample size previously collected at the same agencies) had the lowest ESs of .14 and .15, respectively.

## Discussion

Application of the IISRs to a database of parent ratings of 896 child clients receiving psychotherapy resulted in the creation of brief change-sensitive scales with adequate reliability and larger ESs than scales created for other purposes. Twelve PE-BIMAS items evidenced statistically significant change in at least one of two randomly constructed samples. Although modest, positive change was expected on the basis of the literature review, no a priori reason existed to predict which items would show change, and changes on these 12 items were spread across content domains and included depression, impulsivity, and physical aggression. Nine of these 12 change-sensitive items were cross-validated across both samples.

What makes the PE-BIMAS unique is not its item content per se, but the combination of item content, item phrasing, test instructions, and response format that have been empirically demonstrated to be change sensitive. Lipsey (1998) argued that many evaluations of outcome lack sufficient power to detect intervention effects and that one of the major culprits is dependent measures. As the professionals who typically possess the best training in research design, statistics, and measurement, psychologists use the tools suited for the particular purpose at hand. When asked by stakeholders such as parents, principals, and school boards to evaluate the outcomes of provided interventions, school psychologists would appear well advised to use change-sensitive measures that are sensitive to small and moderately sized treatment effects.

Results from several IISR analyses provided data likely to be of interest to researchers and practitioners. Boys consistently evidenced higher ESs than girls across different change scales; if the client sample were older, the literature suggests that girls may have shown a greater degree of change (Weisz et al., 1998). The content of boys' change items was similar to girls' but also included improvement in additional domains, including academic performance. The academic improvement finding for boys is intriguing because some funding sources, particularly in school settings, may be predicated on the assumption that a link exists between school achievement and mental health variables (cf. Finn, Pannozzo, & Voekl, 1995). Boys' greater improvement may be attributed to the provided interventions as well as to methodological aspects of the study, including the larger sample size of boys, differences in their pretest scores compared to girls, and the sample's composition of elementary school-age children.

The result that a depression item was among those found to be change sensitive is potentially relevant to theories about mechanisms of change in counseling. This finding parallels other IISR studies that have found depression and anxiety items to demonstrate larger ESs after treatment than any other content domain (Meier & Vermeersch, 2007). Meier and Vermeersch (2007) suggest that depression and anxiety may represent a common outcome factor and that alleviation of negative affect may be a ubiquitous effect of all therapeutic interventions. If true, assessments that include negative affect items may represent a universal measure of outcome, a long sought-after criteria in counseling and psychotherapy research and practice (e.g., Frank & Kirk, 1975).

Although IISR scales evidence benefits in outcome assessment compared to traditional tests, conceptual and resource limitations remain. Particularly if test developers cannot obtain a control group, as is often the case in field settings, change evident in a subset of items could result from a number of factors, including practice effects, types of interventions, types of presenting problems, or demand characteristics of the setting. Demand characteristics or socially desirable responding might have been engendered in this study by the use of informed consent that indicated that provided services were being evaluated; future research with the PE-BIMAS should investigate this possibility. Also, the typical test development process involves creating many more items than needed in the final version and then reducing that number during scale administration and item analysis. The reluctance of data providers in settings such as schools and community clinics to complete lengthy questionnaires on more than one occasion will hamper such efforts, often leading to a sample size reduction. Respondents with more severe problems are also less likely to provide information and stay in therapy, again lowering sample sizes and raising questions about the generalizability of results (Meier & Letsch, 2000).

The IISR approach to test construction and evaluation attempts to identify effects at the item level, likely to be small in size, through the use of multiple statistical tests. This runs counter to the statistical philosophy of minimizing the effects of chance through the use of a Bonferroni or similar adjustment (Meier, 2000). One solution to these differing approaches is to consider initial IISR studies as exploratory in nature, using a less stringent $\alpha$. Items identified with an $\alpha$ of .10, as was done in this study, can then be subjected to a cross-validation before any substantive results with those scales are used for reaching conclusions or making decisions during program evaluation, outcome assessment, or theory-building studies. If resources are unavailable for such a replication, then the more stringent Bonferroni-guided scale should be used.

## References

Achenbach, T. M. (1994). Child Behavior Checklist and related instruments. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (pp. 517-549). Hillsdale, NJ: Lawrence Erlbaum.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Conners, C. K. (1994). Conners Rating Scales. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (pp. 550-578). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J., Ambron, S. R., Dornbursch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., et al. (1980). *Toward reform of program evaluation.* San Francisco, CA: Jossey-Bass.

Finn, J. D., Pannozzo, G. M., & Voelkl, K. E. (1995). Disruptive and inattentive-withdrawn behavior and achievement among fourth graders. *Elementary School Journal*, *95*, 421-434.

Frank, A. C., & Kirk, B. A. (1975). Differences in outcome for users and nonusers of university counseling and psychiatric services: A 5-year accountability study. *Journal of Counseling Psychology*, *22*, 252-258.

Hill, C. E., & Lambert, M. J. (2004). Methodological issues in studying psychotherapy processes and outcome. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (5th ed., pp. 84-135). New York: John Wiley.

Kazdin, A. E. (2000). *Psychotherapy for children and adolescents: Directions for research and practice*. New York: Oxford Press.

Lambert, M. J., Hatch, D. R., Kingston, M. D., & Edwards, B. C. (1986). Zung, Beck, and Hamilton Rating Scales as measures of treatment outcome: A meta-analytic comparison. *Journal of Consulting and Clinical Psychology*, *54*, 54-59.

Lewis, J. D., & Magoon, T. M. (1987). Survey of college counseling centers' follow-up practices with former clients. *Professional Psychology: Research and Practice*, *18*(2), 128-133.

Lipsey, M. (1998). Design sensitivity: Statistical power for applied experimental research. In L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 39-68). Thousand Oaks, CA: Sage.

Meier, S. (1997). Nomothetic item selection rules for tests of psychological interventions. *Psychotherapy Research*, *7*, 419-427.

Meier, S. (1998). Evaluating change-based item selection rules. *Measurement and Evaluation in Counseling and Development*, *31*, 15-27.

Meier, S. (2000). Treatment sensitivity of the PE Form of the Social Skills Rating Scales: Implications for test construction procedures. *Measurement and Evaluation in Counseling and Development*, *33*, 144-156.

Meier, S., & Letsch, E. (2000). Data collection issues in an urban community mental health center: What is necessary and sufficient information for outcome assessment? *Professional Psychology: Research and Practice*, *31*, 409-411.

Meier, S. T. (2004). Improving design sensitivity through intervention-sensitive measures. *American Journal of Evaluation*, *25*, 321-334.

Meier, S. T., McDougal, J. L., & Bardos, A. (in press). *The Behavioral Intervention Monitoring Assessment System.* Toronto: Multi-Health Systems.

Meier, S. T., & Vermeersch, D. (2007). *What changes in counseling and psychotherapy?* Manuscript submitted for publication.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Speer, D. C., & Newman, F. L. (1996). Mental health services outcome evaluation. *Clinical Psychology: Science and Practice*, *3*, 105-129.

Stiffman, A. R., Orme, J. G., Evans, D. A., Feldman, R. A., & Keeney, P. A. (1984). A brief measure of children's behavior problems: The Behavior Rating Index for Children. *Measurement and Evaluation in Counseling and Development*, *16*, 83-90.

Vermeersch, D. A., Lambert, M. J., & Burlingame, G. M. (2000). Outcome questionnaire: Item sensitivity to change. *Journal of Personality Assessment*, *74*, 242-261.

Vermeersch, D. A., Whipple, J. L., Lambert, M. J., Hawkins, E. J., Burchfield, C. M., & Okiishi, J. C. (2004). Outcome Questionnaire: Is it sensitive to changes in counseling center clients? *Journal of Counseling Psychology*, *51*, 38-49.

Webster-Stratton, C. (1996). Early onset conduct problems: Does gender make a difference. *Journal of Consulting & Clinical Psychology*, *64*, 540-551.

Weinstock, M., & Meier, S. T. (2003). A comparison of two item selection methodologies for measuring change in university counseling center clients. *Measurement and Evaluation in Counseling & Development*, *36*, 66-75.

Weisz, J. R., Huey, S. J., & Weersing, V. R. (1998). Psychotherapy outcome research with children and adolescents: The state of the art. In T. H. Ollendick & R. J. Prinz (Eds.), *Advances in clinical child psychology* (Vol. 20, pp. 49-91). New York: Plenum.

Weisz, J. R., Weiss, B., Han, S. S., Granger, D. A., & Morton, T. (1995). Effects of psychotherapy with children and adolescents revisited: A meta-analysis of treatment outcome studies. *Psychological Bulletin*, *117*, 450-468.

**Scott T. Meier,** PhD, is a professor and chair of the Department of Counseling, School, and Educational Psychology at the University at Buffalo. His latest book, *Compassionate Data: Measuring Change in Counseling and Psychotherapy*, will be published shortly by Guilford Press.

**James McDougal,** PhD, is an assistant professor in the School Psychology Program in the Counseling and Psychological Services Department at the State University of New York at Oswego. He was formerly the mental health coordinator for the Syracuse City School District, where he provided mental health and behavioral consultation services to over 40 schools and programs.

**Achilles Bardos,** PhD, is a professor at the University of Northern Colorado. He is coauthor of two tests, the General Ability Measure for Adults (GAMA) and the Basic Achievement Skills Inventory (BASI).