Journal of Pediatric Psychology Advance Access published May 31, 2009

Editorial: An Author's Checklist for Measure Development and Validation Manuscripts

Grayson N. Holmbeck and Katie A. Devine
Loyola University Chicago

A recent special issue of the *Journal of Pediatric Psychology* included papers focused on evidence-based assessment across several broad domains of assessment in pediatric psychology (e.g., adherence, pediatric pain, and quality of life). In one of these papers, Holmbeck et al. (2008) reviewed strengths and limitations of existing measures of psychosocial adjustment and psychopathology, concluding that many measures lacked supporting psychometric data (e.g., basic indices of reliability and validity) that would permit a complete evaluation of these measures. Given that measure development and validation papers are frequently published in *JPP* (Brown, 2007), it is important that we attend to guiding psychometric principles when developing and disseminating data on new measures to be employed with pediatric populations (Nunnally & Bernstein, 1994). Thus, the purpose of this paper is to present and describe a checklist for authors to use when submitting measure development papers to *JPP*. This checklist is included in the Appendix and is also included at the following link on the *JPP* website:

http://www.oxfordjournals.org/our_journals/jpepsy/for_authors/measure%20development%20checklist.pdf

Findings presented by Holmbeck et al. (2008) indicated that 34 of the 37 measures reviewed met pre-established ''evidence-based assessment'' (EBA) criteria for ''well-established'' measures (Cohen et al., 2008). To be considered ''well-established,'' a measure had to have been presented in at least two peer-reviewed journal articles by different investigatory teams, have demonstrated adequate levels of reliability and validity, and be accompanied by supporting information (e.g., a measure manual). Although most measures that we reviewed met these criteria, we also found that most of the 34 ''well-established'' measures were hampered by at least one major psychometric flaw and/or lacked important psychometric data. We concluded that a more fine-grained EBA classification system is needed.

One important distinction in this literature relates to differences between empirically supported assessment and evidence-based assessment. This type of distinction was first discussed in the literature on clinical interventions (e.g., Spring, 2007). An empirically supported intervention is one that has demonstrated efficacy in randomized clinical trials or clinic-based effectiveness trials. An evidence-based intervention has empirical support in the manner just described, but also ''integrates research evidence, clinical expertise, and patient preferences and characteristics . . . empirically-supported treatments (ESTs) are an important component of evidence-based practice (EBP), but EBP cannot be reduced to ESTs'' (Spring, 2007, p.611). Applying these terms to the field of assessment and measure development efforts, an empirically supported assessment measure would be one that demonstrates satisfactory psychometric characteristics, broadly defined. To be evidence based, the measure should also demonstrate utility in clinical settings, be useful in making diagnoses, be sensitive to treatment effects, and/or provide incremental validity above and beyond other similar measures. Although papers in the special issue of *JPP* frequently referred to ''evidence-based assessment'' (Cohen et al., 2009), the articles included in the issue tended to evaluate the degree to which the measures were empirically-supported rather than evidence based. To be ''evidence-based,'' our reviews would have needed to *integrate* an evaluation of clinical utility, diagnostic utility, and treatment sensitivity with the empirical psychometric data presented in each review. As noted, the published reviews were more likely to focus on the latter rather than on the former.

As suggested by Mash and Hunsley (2005), detailed EBA profiles would provide a complete evaluation of

All correspondence concerning this article should be addressed to Grayson N. Holmbeck, Loyola University Chicago, Department of Psychology, 6525 N. Sheridan Road, Chicago, IL 60626. E-mail: gholmbe@luc.edu

Journal of Pediatric Psychology pp. 1–6, 2009
doi:10.1093/jpepsy/jsp046

evidence across each of several psychometric and clinically relevant dimensions, including: (a) internal consistency, (b) test–retest reliability, (c) the availability of normative data, (d) content validity, (e) construct validity, (f) convergent and discriminant validity, (g) criterion-related validity, (h) incremental validity, (i) clinical utility, (j) diagnostic utility, and (k) treatment sensitivity. The focus on incremental validity and clinical and diagnostic utility raises the bar from a focus on ''empirical support'' (i.e., where the focus would tend to be primarily on psychometric data) to a broad focus on the ''evidence base'' for a measure. In developing the checklist that is the focus of this article, we attempted to provide a list of criteria relevant to establishing the evidence base (and not just empirical support) for a measure. In addition to shifting the focus from providing ''empirical support'' for a measure to providing an ''evidence base'' for our instruments, a checklist for measure development papers would permit *JPP* reviewers to evaluate such papers in the same way that reviewers of randomized clinical trials make use of the Consolidated Standards of Reporting Trials (CONSORT) checklist and flowchart (Altman et al., 2001). The CONSORT checklist contains reporting standards with respect to methodological features of and the manner in which results are reported in clinical trials. Moreover, authors are required to provide a flowchart that describes details of sample recruitment and attrition during the course of the study.

Thus, a checklist for measure development papers would serve two interrelated purposes: (a) it would provide guidance to authors as they embark on the measure development process and would provide a list of criteria authors can use as they develop an evidence base for their measures, and (b) it would begin to standardize the manner in which psychometric and other assessment-related data are presented in measure development papers for this journal. Before providing a more detailed overview of the checklist, it is important to note that this checklist is rather exhaustive (see Appendix). As such, it represents what would ''ideally'' be expected for a measure development or validation manuscript rather than minimal criteria for such papers. No one paper can provide a complete evaluation of all important psychometric and clinically relevant dimensions that will establish once-and-for-all the evidence base for a measure.

Instrument refinement is part of a measure development process that gradually builds an evidence base for a scale (see Smith & McCarthy, 1995, for suggestions on measure refinement). Indeed, the validation of any measure is a cumulative process that occurs across many different types of research studies and across research programs.

Overview of Checklist for Measure Development Papers

As can be seen in the Appendix, the first, and perhaps most important criterion, focuses on the degree to which the author has established a scientific need for the instrument. This is a fundamentally important criterion that should be included in all measure development papers. How does this measure make a contribution to the literature and/or clinical practice above and beyond other previously developed measures? How will the measure be used and by whom?

With respect to the scientific necessity of the measure, it is worth discussing one type of manuscript that is often submitted to this journal. Many authors seek to employ a given measure with a new population that differs from the population that was the basis for the original measure development research. Given that the number of ''new populations'' to which a measure can be applied is infinite, these types of papers benefit greatly from a clearly articulated rationale for why it is of interest to employ the measure with this particular ''new population.'' For example, one might discuss how the construct of interest is relevant to this population and whether there are important differences in how the construct would be perceived in this population as compared to how it would be perceived in other populations. Simply stating that this construct has never been assessed in a given population is not a sufficient justification for applying a measure to a new population.

Once the author has determined that there is a need for this measure either for research and/or clinical purposes, one typically attends to issues of content validity *prior to* actually developing the measure or generating items (Haynes, Nelson, & Blaine, 1999; Haynes, Richard, & Kubany, 1995). Although it is often tempting to begin developing a measure based on one's own knowledge of the construct of interest or the urgent need to develop a measure for use in a larger research project, the ''content validity'' stage is one of the most important parts of the measure development process. Indeed, we often receive submitted manuscripts where it is clear that items were generated by a research team that was not necessarily made up of experts with respect to the construct of interest. It is important to take the time to properly define your construct and specify dimensions that underlie the construct. Item generation can be based on a variety of factors and strategies (see Appendix), including a review of the larger research literature and consultation with experts and relevant target populations. During the item generation stage, it is important to maintain roughly equivalent

numbers of items across dimensions and to generate more items than are necessary so that items that do not function in an appropriate manner psychometrically can be eliminated at later stages of the measure development process (Nunnally & Bernstein, 1994). If one starts with too few items, one may end up with small subscales of questionable psychometric quality. The reading level of all items should be assessed and measure instructions and the response scale need to be developed (see Clark & Watson, 1995, and Comrey, 1988, for information on how to generate appropriate items). Additionally, it is important to determine whether the measure would be appropriate across multiple developmental levels and for different ethnic groups (Frick, 2000). After an item pool has been generated, it is useful to again consult with experts and/or members of the target population to assess for item relevance and wording ambiguities.

Once content validity has been ''built in'' to the measure during this important initial stage of measure development, the investigator can begin to gather data on reliability indices and conduct item analyses (see Appendix for details). Problematic items can be dropped during this stage and the hypothesized dimensions of the construct can be evaluated via confirmatory factor analyses (which may result in more problematic items being dropped from the scale). If one employs an exploratory factor analysis, several difficulties often emerge. One may retain too many factors, thus yielding subscales with small numbers of items or one may ''accept'' an unsatisfactory solution where many items load significantly on more than one subscale.

After adequate reliability and factorial integrity have been established, the investigator can develop a plan to test the validity of the measure. This process may involve multiple studies with different types of validation samples (e.g., one may want to compare scores on the measure across clinical and nonclinical samples). The measure should exhibit high correlations with other measures that tap similar constructs and it should be less highly correlated with measures that assess different constructs. Moreover, one can expect that scores on the measure will be associated with other behaviors assessed concurrently or prospectively. A difficulty that often arises at this stage of the measure development process is that the researcher will employ self-report methods for both the measure of interest ''and'' for the validity indices, making it impossible to rule out common method variance interpretations for the findings. In other words, such a study has limited potential to evaluate the validity of a measure.

Finally, the investigator may be interested in documenting the utility of the measure in assessing responsiveness to treatment or in making diagnostic decisions. Measuring responsiveness to treatment requires additional considerations, such as whether the scale can be administered repeatedly, whether it is sensitive to change over time, and how much change reflects meaningful differences in a person's functioning (Kazdin, 2005). One may also be interested in assessing the degree to which the measure is predictive of outcomes above and beyond other existing measures and the degree to which the measure is cost effective in a clinical setting (i.e., is the information provided worth the time allotted for its administration and scoring?). If relevant, one may also be interested in employing appropriate procedures for translating the measure into languages other than English (see Appendix for details regarding the translation process).

Conclusions

In this paper, we have described a checklist for those who seek to submit measure development and measure validation papers to *JPP*. In attempting to promote evidence-based assessment, we have highlighted the importance of attending to treatment sensitivity and diagnostic and clinical utility when developing a new measure. We hope that the checklist provided will not only be useful for authors but for reviewers as well. Again, we note that this checklist represents what would ideally be expected for measure development over multiple studies rather than minimal criteria for a single study.

Funding

A research grant from the National Institute of Child Health and Human Development (R01-HD048629).

Conflicts of interest: None declared.
Received, revisions received and accepted April 30, 2009

Appendix

Criteria and Checklist for Measure Development Papers

☐ 1. Establishes Scientific Need for the Instrument

- ☐ a. Reviews research and/or clinical practices to establish need for the instrument
- ☐ b. Specifies the new contribution of the measure relative to previous research

☐ 2. Attends to Content Validity During Initial Measure Development (based on Clark & Watson, 1995;

Haynes, Richard, & Kubany, 1995; Haynes, Nelson & Blaine, 1999)

- ☐ a. Defines the construct
  - ☐ i. Reviews theory underlying the construct
  - ☐ ii. Specifies what will be included and excluded in the measure
  - ☐ iii. Specifies facets or dimensions of construct
- ☐ b. Specifies contexts/situations for the measure
  - ☐ i. Specifies setting for completion of measure
- ☐ c. Specifies intended function of the measure
  - ☐ i. Specifies purpose of measure
  - ☐ ii. Specifies target population
  - ☐ iii. Specifies appropriate age range
  - ☐ iv. Determines if appropriate for multiple developmental levels and ethnic groups
- ☐ d. Selects and generates items based on:
  - ☐ i. Clinical experience
  - ☐ ii. Relevant theories
  - ☐ iii. Empirical literature
  - ☐ iv. Rational deduction
  - ☐ v. Related Instruments
  - ☐ vi. Consultation with experts
  - ☐ vii. Focus groups with target population
- ☐ e. During item generation, matches items to facets/dimensions
  - ☐ i. Includes appropriate numbers of items for each dimension
  - ☐ ii. Attends to test length (generates an appropriate number of items given the setting in which it will be used, generates enough items to allow for some items to be dropped during the test refinement process, generates enough items to assess the construct)
- ☐ f. Conducts qualitative item analysis (relevance of each item, wording of items, check for redundancy across items)
- ☐ g. Addresses literacy and reading level issues for the target population
- ☐ h. Determines response format and scoring method
  - ☐ i. Selects response format (e.g., Likert, etc.)
  - ☐ ii. Attempts to reduce impact of response sets by not wording all items in same direction
  - ☐ iii. Scoring method is explained
- ☐ i. Develops appropriate instructions for measure (including time frame; e.g., ''During the past two weeks . . .'')
- ☐ j. Has experts review the initial version of the instrument
- ☐ k. Has members of target population review initial version of the instrument
- ☐ l. After refinement of measure:
  - ☐ i. Additional item analysis
  - ☐ ii. Additional review by experts
  - ☐ iii. Additional review by members of target population
- ☐ m. Conducts pilot testing of measure

☐ 3. Evaluation of Reliability

- ☐ a. Evaluates internal consistency (subscales, full scale)
- ☐ b. Evaluates temporal stability (test–retest)
- ☐ c. Uses generalizability theory in assessing reliability
- ☐ d. Cross-validates reliability estimates

☐ 4. Develops Norms for the Measure

- ☐ a. Develops norms for different relevant populations

☐ 5. Quantitative Item Analysis

- ☐ a. Examines whether items discriminate between relevant groups
- ☐ b. Includes corrected item-to-total correlations
- ☐ c. Includes average correlations between individual items and all other items
- ☐ d. Evaluates distributions of items and eliminates items with inadequate distributions
- ☐ e. Evaluates items using Item Response Theory (particularly if it is a measure that assesses abilities or skills)
  - ☐ i. Examines item characteristic curves (see Nunnally & Bernstein, 1994)
  - ☐ ii. Examines unidimensionality of items, the appropriateness of using a total summary score, and differential item functioning using Rasch analysis (Tennant, McKenna, & Hagell, 2004; Tesio, 2003)

☐ 6. Conducts Factor Analyses

- ☐ a. Evaluates factor structure of measure via exploratory factor analyses/principal components analyses

- ☐ b. Confirms hypothesized factor structure of measure via confirmatory factor analyses

☐ 7. Evaluation of Validity

- ☐ a. Clearly articulates plan for assessing validity
- ☐ b. Includes a priori hypotheses for major analyses
- ☐ c. Evaluates overall construct validity of measure (which involves a general evaluation of all validity evidence for the measure)
- ☐ d. Evaluates convergent validity, which is the degree of convergence between the target measure and other instruments purporting to measure the same construct
- ☐ e. Evaluates discriminant validity, which is the degree to which the target measure is not associated with other measures that assess different constructs
- ☐ f. Evaluates criterion-related validity, which is the degree to which scores on the target measure are associated with measures of non-test behaviors (includes concurrent and predictive validity)
- ☐ g. Cross-validates validity estimates

☐ 8. Evaluates Diagnostic Utility, Clinical Utility, and Cost-Effectiveness (based on Haynes et al., 1999)

- ☐ a. Evaluates degree of treatment utility
  - ☐ i. Is the measure sensitive to change?
  - ☐ ii. Can it be used repeatedly over the course of treatment and does it reflect improvement or worsening of symptoms? (see Kazdin, 2005)
- ☐ b. Evaluates degree of diagnostic utility (see Bossuyt et al., 2003)
  - ☐ i. Includes estimates of diagnostic accuracy (sensitivity, specificity, positive and negative predictive power)
- ☐ c. Evaluates degree of incremental validity (does the measure add value in clinical judgment above and beyond other measures?)
- ☐ d. Evaluates measure's cost-effectiveness

☐ 9. Translates Measure into Other Languages

- ☐ a. Semantic equivalence: Translation by a native speaker and back-translation by an independent native speaker. Region-specific language should be used when possible
- ☐ b. Content equivalence: Native language speaker has reviewed content of items for appropriateness and equivalence
- ☐ c. Technical equivalence: All language versions contain the same item and scale formatting

References

Altman, D. G., Schulz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., et al. (2001). The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. *Annals of Internal Medicine, 134*, 663–694.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glaszion, P. P., Irwig, L. M., et al. (2003). The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. *Annals of Internal Medicine, 138*, W1–W12.

Brown, R. Y. (2007). Journal of Pediatric Psychology (JPP), 2003–-2007; Editor's Vale Dictum. *Journal of Pediatric Psychology, 32*, 1165–1178.

Cohen, L. L., La Greca, A. M., Blount, R. L., Kazak, A. E., Holmbeck, G. N., & Lemanek, K. L. (2008). Introduction to Special Issue: Evidence-based assessment in pediatric psychology. *Journal of Pediatric Psychology, 33*, 911–915.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment,* 7, 309–319.

Comrey, A. L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology, 56*, 754–761.

Frick, P. J. (2000). Laboratory and performance-based measures of childhood disorders: Introduction to the special section. *Journal of Clinical Child Psychology, 29*, 475–478.

Haynes, S. N., Nelson, K., & Blaine, D. D. (1999). Psychometric issues in assessment research. In P. C. Kendall, J. N. Butcher, & G. N. Holmbeck (Eds.), *Handbook of research methods in clinical psychology* (2nd ed., pp. 125–154). New York: Wiley.

Haynes, S. N., Richard, D. C. S, & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment,* 7, 238–247.

Holmbeck, G. N., Welborn Thill, A., Bachanas, P., Garber, J., Miller, K. B., Abad, M., et al. (2008). Evidence-based assessment in pediatric psychology: Measures of psychosocial adjustment and psychopathology. *Journal of Pediatric Psychology, 33*, 958–980.

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

Mash, E. J., & Hunsley, J. (2005). Evidence-based assessment of child and adolescent disorders: Issues and challenges. *Journal of Clinical Child and Adolescent Psychology, 34*, 362–379.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill..

Smith, G. T., & McCarthy, D. M. (1995). Methodological considerations in the refinement of clinical assessment instruments. *Psychological Assessment, 7*, 300–308.

Spring, B. (2007). Evidence-based practice in clinical psychology: What it is, why it matters, what you need to know. *Journal of Clinical Psychology, 63*, 611–631.

Tennant, A., McKenna, S. P., & Hagell, P. (2005). Application of Rasch analysis in the development and application of quality of life instruments. *Value in Health, 7*, S22–S26.

Tesio, L. (2003). Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. *Journal of Rehabilitation Medicine, 35*, 105–115.