Published in final edited form as:

Educ Psychol Meas. 2006 August ; 66(4): 687–700. doi:10.1177/0013164405282467.

The Development and Evaluation of Procedures to Assess Child Self-report Item Validity

Michael E. Woolley,
University of Michigan, Ann Arbor

Gary L. Bowen, and
University of North Carolina, Chapel Hill

Natasha K. Bowen
University of North Carolina, Chapel Hill

Abstract

Cognitive pretesting (CP) is an interview methodology for pretesting the validity of items during the development of self-report instruments. The present research evaluates a systematic approach to the analysis of CP data. Materials and procedures were developed to rate self-report item performance with CP interview text data. Five raters were trained in the application of that system. Estimates of inter-rater reliability found acceptable to substantial levels of inter-rater agreement. Results from the present study suggest that excellent inter-rater reliability can be achieved in the evaluation of CP data. Guidelines for systematically rating the qualitative data collected using CP methods are provided. Future research should focus on empirical demonstrations of how such rating procedures can lead to improvements in self-report instruments.

Cognitive pretesting (CP) is one of many cognitive approaches to assessing the validity of self-report items. The CP methodology is to interview individuals while they read and respond to self-report items, in order to collect data about item comprehension and response (DeMaio & Rothgeb, 1996; Jobe & Mingay, 1989). Cognitive methods to assess the validity of self-report items have been applied with adults since the 1980s (Jabine, Straf, Tanur, & Tourangeau, 1984), and the CP methodology has more recently been adapted for use with children (Bowen, Bowen, & Woolley, 2004; Rebok et al., 2001). This methodology seems especially helpful in examining the developmental validity of self-report items for children (Woolley, Bowen, & Bowen, 2004).

Researchers utilizing CP have asserted its utility for improving item performance (McKay & de la Puente, 1996; Sirken et al., 1999). Recently, however, some authors have expressed concern about the lack of systematic and consistent procedures for CP data collection and analysis (Foddy, 1998; Willis, DeMaio, & Harris-Kojetin, 1999). The present research reports on the development of a systematic approach to coding item validity performance by utilizing CP data. This system was developed and evaluated during the construction of the Elementary School Success Profile (ESSP) child questionnaire (Bowen et al., 2004), a self-report instrument for use with children in middle childhood. The CP data coding system reported here includes two components: a validity codebook detailing item performance rating criteria and procedures, and rater training in the application of those criteria and procedures. The efficacy of these procedures is evaluated by assessing the inter-rater reliability achieved when five trained raters apply the rating system.

Cognitive Pretesting

The complex task of constructing valid and reliable self-report instruments can be further complicated by cognitive developmental issues, when constructing such instruments for use with children (Woolley et al., 2004). Standard scale-development procedures, such as collecting pilot data and utilizing statistical approaches to assess self-report instrument performance, can identify items that perform poorly in relation to other items within a scale (DeVellis, 2003). However, utilizing cognitive approaches to pretest self-report items can provide richer data about individual item performance. Such data can be used to rate item performance by providing information about how respondents interpret an item, about what respondents think about while processing the item, and about the rationale respondents use to choose an answer option for that item (DeMaio & Rothgeb, 1996; Foddy, 1998). By providing data about *how* respondents process an item, cognitive approaches can also inform the modification of items to improve performance.

Cognitive pretesting is one of many approaches, known collectively as *cognitive methods* (Forsyth & Lessler, 1991), for collecting data about self-report items directly from respondents. Essentially, the technique involves interviewing a child while he or she reads, interprets, and responds to a self-report item. Bowen et al. (2004) describe four steps in a CP interview procedure for children: (a) ask the child to read the question out loud; (b) ask the child what the question means, or what the question is asking him or her or is trying to find out; (c) ask the child to read the answer options and to choose an answer; and (d) ask the child to explain why he or she chose that answer.

Two rounds of CP were completed on the ESSP child questionnaire before the research reported here began. The third round of CP provided the opportunity to apply and test the newly developed CP data analysis materials and the procedures reported here. For more extensive discussions of the role of CP in the development of the ESSP, see Bowen et al. (2004) and Woolley et al. (2004).

Inter-rater Reliability

The evaluation of the item validity rating system developed for this study was accomplished by estimating and interpreting the level of inter-rater reliability attained in the rating of CP data. Therefore, four inter-rater reliability issues central to the current research will be discussed: (a) estimating inter-rater reliability coefficients, (b) interpreting inter-rater reliability coefficients, (c) rater confidence, and (d) rater training.

Estimating Inter-rater Reliability Coefficients

There are multiple approaches and formulas utilized to estimate inter-rater reliability. Examples of these strategies include: (a) Cohen's kappa (a chance-corrected measure of rater agreement); (b) Cronbach's alpha coefficient (typically applied to scales, but applicable to multiple raters); (c) the Spearman-Brown formula; (d) the Pearson product-moment correlation coefficient; and (e) intraclass correlation coefficients (a family of correlation formulas applicable to various reliability conditions) (Dunn, 1989; McDonald, 1999; Natsuti & Pecora, 1993; Shrout & Fleiss, 1979). Under certain reliability conditions, various formulas can result in similar coefficients; under other conditions, various formulas can result in divergent results (Cronbach, 1990; Dunn, 1989; Shrout & Fleiss, 1979).

Multiple approaches to estimate inter-rater reliability are utilized because of the complex nature of inter-rater reliability—various conditions in which inter-rater reliability are estimated call for different formulas. Choosing an inappropriate formula can lead to over- or under-estimation of the reliability of multiple raters in a given situation. The key is to

choose a formula that fits the reliability conditions being studied, including the nature of the raters, the data, and the intended interpretation of the results.

Shrout and Fleiss (1979) present a system for choosing the appropriate ICC formula for the inter-rater reliability conditions under study. Choosing the ICC appropriate for specific conditions can be accomplished by answering three questions: Is a one-way or a two-way analysis of variance appropriate for the condition? Is the unit of analysis one rating or the mean of multiple ratings? And finally, are differences in the mean ratings of raters significant?

The answer to the first question denotes a condition of the raters. If each datum is rated by a different set of raters, then a one-way analysis of variance is indicated. However, if the same set of raters rate all data, the calculation of the ICC requires a two-way analysis-of-variance model. The second question is a function of whether the calculation sought is an estimate of the reliability of any one rater or the reliability of the mean rating of all raters.

The answer to question three is dependent upon whether the raters are considered a random or a fixed effect. This decision has important ramifications both for the formula used and for the interpretation of the results. When treating raters as a fixed effect, the appropriate ICC estimates rater *consistency* and interpretation of the results is limited to the current raters. When estimating rater consistency, calibration differences between raters are removed from the total error variance in the denominator, as seen in the formula for the two-way mixed effect ICC:

$$ICC(2, k) = \frac{BMS - EMS}{BMS + (JMS - EMS)/n}$$

When treating raters as a random effect, on the other hand, the appropriate ICC estimates rater *agreement* and allows generalizability to other, similar sets of raters. When estimating rater agreement, calibration differences between raters are included in the total error variance, as seen in the formula for the two-way random effect ICC:

$$ICC(3, k) = \frac{BMS - EMS}{BMS}$$

Interpreting Inter-rater Reliability Coefficients

Setting benchmarks for inter-rater reliability coefficients is complex. Authors who have presented tables to guide the interpretation of inter-rater reliability coefficients assert that such tables suffer from an element of subjectivity, due to the varying nature of inter-rater reliability situations (Dunn, 1989; Shrout, 1998). After recognizing that element of subjectivity, Shrout proposes the following inter-rater reliability benchmarks, which will be applied in the current study: 0.00 to 0.10 (*virtually none*), 0.11 to 0.40 (*slight*), 0.41 to 0.60 (*fair*), 0.61 to 0.80 (*acceptable*), and 0.81 to 1.0 (*substantial*).

Rater Confidence

The self-reported confidence of raters has been shown to be an important variable when interpreting inter-rater reliability coefficients (Fox, Bizman, Hoffman, & Oren, 1995; Kavanaugh, 1989). Assessing the confidence of raters along with inter-rater reliability allows a more complete assessment of a rating system, of the raters applying that system, and of the data rated (B. D. Goldman, personal communication, June 11, 2003). For example, Kavanaugh found that rater confidence was positively correlated with rater

accuracy, indicating that raters are able to effectively self-evaluate the confidence they have in their own ratings. Rater confidence has also been shown to be an indicator of how difficult the data are to rate. For example, Fox et al. found that when the phenomena rated showed higher levels of variability, raters were less confident. In addition, there is a curvilinear relationship between item quality and rater confidence: raters are less confident about data rated in the middle of a rating scale, while they are more confident of high or low ratings.

Rater Training

Rater training is an important consideration in the pursuit of reliable, confident, and accurate raters (R. F. DeVellis, personal communication, April 3, 2003). For example, Weigle (1998) found that experienced raters of text data were much more consistent than inexperienced raters. However, after the raters were provided with training, all raters showed improved consistency, and no differences could be seen between experienced and inexperienced raters. Training also reduced the number of extreme scores previously obtained from inexperienced raters. Dyrborg et al. (2000) likewise found that raters who practiced with a rating system achieved a higher level of consistency in their ratings.

Woehr and Huffcutt (1994) conducted a meta-analysis of four rater training approaches and found that all approaches increased accuracy, decreased error, and increased rater observation skills. However, the most effective approach was “frame-of-reference training,” which includes (a) clear standards for rating, (b) examples of ratings, and (c) a process whereby raters “share and use common conceptualizations of performance” (p. 192).

Methods

The methods and procedures employed in this study include a codebook, rater procedures, and rater training. These materials, procedures, and activities were designed to provide a system for rating the validity of child self-report item performance in a reliable and replicable manner. Three hierarchical validity performance criteria were at the core of the rating system. The codebook provided operational definitions for all three performance criteria, applied to 29 questionnaire items pretested in the study. Five raters were utilized to rate item validity performance, and all received the rater training in a group setting. In the subsections that follow, we describe the dataset used in the study, detail the codebook and rater procedures, describe the raters, outline the rater training sessions, and describe the data collection process.

The CP Dataset

The CP interview data utilized in the current research were collected during the third round of CP of the ESSP child questionnaire. The third-round CP interviews were conducted with children in an after-school program at an elementary school in North Carolina. Informed consent was obtained from a parent/guardian, and assent was obtained from each child. The interview procedures were audiotaped, and the child responses to the second, third, and fourth CP interview questions were transcribed. Each child-by-item set of data therefore included (a) a child's interpretation of the item, (b) a chosen answer option, and (c) an explanation for the answer option chosen.

CP data were collected on all 81 child items; however, a subset of 29 of the items was analyzed using newly developed systematic materials and methods. These 29 items represent five scales within the ESSP child questionnaire. These scales were chosen because they were expected to be the most likely to reveal validity problems in light of the findings in the first two rounds of pretesting.

This data sample was obtained from 15 children who ranged in age from 8 to 11, and included 5 third-graders, 4 fourth-graders, and 6 fifth-graders. Six of the participants were girls and 9 were boys. Eleven of the students were European American, 2 were African American, 1 was Hispanic, and 1 was of mixed ethnicity. The transcribed child responses from round three constitute five out of the six child-by-item interview datasets used in the current research.

All child questionnaire items had already undergone a rigorous development process, including two previous rounds of CP (Bowen et al., 2004); consequently, a high percentage of items were expected to perform well, which reduced the variance in the item performance ratings. Therefore, one set of child-by-item data was added to the dataset, to increase the variance in the ratings. The sixth set of child-by-item data included data from the second round of CP; some responses had low validity and were used verbatim, while some were modified to have low validity. Therefore, the dataset utilized in this study included interview responses from 6 children, and 29 self-report items (constituting five subscales), for a total of 174 child-by-item CP interview responses.

The Codebook

The codebook included two key components: three validity performance criteria, and CP data examples illustrating the application of those criteria. Each codebook page also included the item and the response options (see Figure 1).

The first rating criterion involved the *concept* the item was designed to measure. The concept of an item was defined as the aspect of the construct targeted by the scale that a specific item measured. For an interviewee to demonstrate comprehension of the item concept, his or her responses to the CP interview questions should have indicated an interpretation of the item that matched the defined intention of the item. The concept criterion was defined in the codebook as follows: “Concept—Does the child comprehend *what* the question is asking about?”

The second criterion, *coherence*, defined what constituted retrieval and articulation of memory relevant to the concept targeted by the item. The intent of this criterion was to assess what a respondent thought about while processing and responding to the item. CP interview responses appropriate to assessing coherence included a description of experiences, anecdotes, examples, occurrences, interactions, or relationships relevant to the concept targeted by the item. Therefore, the codebook guidelines for assessing coherence anticipated and described respondent life experiences that were considered relevant and acceptable in response to the item concept. The coherence criterion was defined in the codebook as follows: “Coherence—Does the child describe thoughts, feelings, interpersonal interactions, specific situations, and/or a general pattern of events that reflect who, where, and when the question is asking about?”

The third criterion, *congruence*, established guidelines for assessing the relationship between the respondent’s answer choice and everything he or she had said about the item during the interview. Congruence between the chosen answer option and the respondent’s experience with the item concept is an essential aspect of the final determination of valid item performance. The congruence criterion was defined in the codebook as follows: “Congruence—Does the child’s coherent description or explanation reflect the answer option chosen?”

These three rating criteria were evaluated hierarchically; higher ratings required meeting criteria for all lower ratings. Therefore, the rating scale for item performance was an ordinal 4-point scale ranging from 0 to 3. A score of 0 represented failure of the item, while 3

represented valid response. Rating scale anchors were defined, to qualitatively distinguish rating levels. Anchoring the rating scale leads to more consistent interpretation and application by raters and therefore makes the scale less subject to calibration differences.

Finally, each codebook page included two examples of CP interview data, with indications in the text where evidence of the criteria may be found. These examples were not actual data but were constructed to read like child responses. Raters were also provided with guidelines and procedures detailing the application of the codebook. These materials were all presented to the raters during the two training sessions described in later sections.

Confidence Ratings

In light of the research findings on rater confidence, discussed previously, a self-reported measure of rater confidence was added to the current study. Raters reported their confidence in individual ratings utilizing a 4-point Likert scale ranging from 1 (*low confidence*) to 4 (*high confidence*). A mean self-reported confidence rating of 3.0 or higher by a rater was set as the benchmark for concluding a rater experienced a substantial level of confidence when applying the rating system.

Raters

Five raters were included in the study, in order to ensure the statistical power necessary for a robust evaluation of inter-rater reliability without threatening feasibility or generalizability. The raters were all females ranging in age from 20 to 27, including two social work bachelor's degree students, two social work master's degree students, and one social work doctoral student. Two raters were African American, two were European American, and one was Hispanic/Latina.

Rater Training

The rater training approach utilized in the current study fits the format of frame-of-reference training, described previously. This format includes clear rating standards and examples of data ratings (both included in the codebook) as well as the opportunity for raters to practice rating data in a group process (the training sessions detailed in the next section). The training process included two 3-hour sessions, one week apart.

Training session one—The goals of the first rater training session were to familiarize the raters with the child questionnaire and CP procedures, and to introduce the raters to the codebook and other rating materials. The outline of the first training session included six activities: (a) an overview of the ESSP, (b) an overview of self-report and validity, (c) an overview of CP, (d) applied CP experience, (e) an introduction to the codebook and rating system, and (f) rater feedback about the rating system.

An active group process makes rater training more effective; therefore, group interaction was encouraged. An overview of the self-report process and of the concept of validity—the central issue in the overall rating process—was included, to provide context for the raters. After the CP procedures were described, raters took turns completing CP interviews on four questionnaire items. Each rater took a turn as an interviewer, as an interviewee, and then a data transcriber. This CP experience was intended to give raters a firsthand understanding of the process used to collect the data they were going to rate.

Next, raters were introduced to the codebook, rater guidelines, and procedures. Raters then rated the transcribed CP interview responses collected during the CP experience. Group discussion after each set of four ratings allowed raters to share input and to compare their own ratings with others'. In some cases, varying ratings of identical data emerged, and

issues were identified related to the rating criteria, the guidelines, and the procedures, which then informed modifications to the rating materials and procedures.

Training session two—The second training session included: (a) a review of the modifications to the rating system and procedures made as the result of feedback from training session one, (b) an overview of the rater confidence rating scale, (c) assignment of rater IDs, (d) an overview of rater practice procedures, (e) rating practice and group discussions, (f) acquiring rater consent, and (g) an overview of the rater packets.

The goals of this session were to practice rating data and to group-process ratings that were not reliable across raters. The raters were given identical data sheets to rate and were asked to indicate a confidence level for each rating. Meanwhile, previously completed data sheets were analyzed, and data that were rated inconsistently were identified. A group discussion was held about any data showing poor inter-rater reliability. The discussions focused on arriving at a group consensus on the appropriate rating for those data, according to the codebook.

The practice and discussion activities continued through seven rounds, until substantial rater reliability was achieved. The reliability coefficients for the seven rounds of practice were as follows: 0.62, -0.04, 0.53, 0.93, 0.90, 0.86, and 0.87. Confidence ratings similarly improved throughout the seven rounds.

Rater Data Collection

At the end of the second training session, raters were each given a rater packet. The packets included data rating sheets for all 29 ESSP child questionnaire items, organized by scale, as well as rater confidence feedback sheets and a form requesting demographic information from each rater. Raters were instructed not to discuss the rating process with one another, but to e-mail or phone the researchers if they had questions. All raters completed their packets without contacting the researchers, and all rater packets were returned within one week.

Analysis Procedures

Five raters rated the validity of 174 child-by-item CP interview response data, utilizing the codebook and rating materials. The ratings were utilized to assess the inter-rater reliability of the CP rating materials and the procedures employed. In order to choose the correct ICC for the reliability conditions of the current study, the specific reliability conditions were identified. These conditions included: (a) all five raters rated all the data; (b) the focus of the analysis was the reliability of the mean rating of all five raters; (c) a goal of the study was to generalize the reliability of the rating system to other applications of the CP methodology; and (d) a finding of substantial rater agreement—as opposed to consistency—on item validity performance would be a more rigorous demonstration of the utility of the rating system.

Utilizing those conditions to answer the three key questions defined by Shrout and Fleiss (1979) indicated that a two-way random effects intraclass correlation coefficient was required. The two-way random effects ICC is abbreviated the ICC (2, 5)—with the 2 indicating the reliability conditions defined by the three critical questions, and the 5 indicating the number of raters utilized. Raters also self-reported their confidence in each CP datum rating made; those ratings were analyzed by examining descriptive statistics of the ratings. All analyses were completed utilizing SPSS 10.0.5 (2000).

Results

Table 1 details the descriptive statistics for the item performance ratings made by each rater. The results showed that all raters applied the entire range of the rating scale, with no out-of-range values. The ratings of the five raters had similar means. The standard deviations were also similar, spanning approximately one unit on the rating scale. The means of the ratings revealed a negatively skewed distribution with a significant number of child-by-item data rated as 3.0, which was anticipated, as was described earlier.

Table 2 presents the inter-rater reliability coefficients and the confidence intervals for each coefficient. Inter-rater reliability coefficients were estimated utilizing the ICC (2, 5) formula. The reliability coefficient that estimated rater agreement for all 29 items was .85, indicating substantial inter-rater agreement. In addition, the inter-rater reliability coefficients for rater agreement for the five scales ranged from acceptable (.63) to substantial (.92).

Rater confidence was assessed using descriptive statistics of the confidence ratings reported by the five raters. These statistics are detailed in Table 3. These results reveal that rater confidence ratings ranged from a mean of 3.14 to 3.87, on a scale from 1 (*low confidence*) to 4 (*high confidence*). As described previously, a mean rater confidence of 3.0 or above was set a priori as representing substantial confidence on the part of raters in their application of the rating system to the CP text data. That benchmark was exceeded for all raters. In fact, only seven confidence ratings fell below 3.0 for all five raters, and none of the five raters reported a confidence rating less than 2.0 for any item performance rating.

Discussion

The intent of the current research was to develop and evaluate a systematic and replicable approach for the analysis of CP data. The approach reported here included an item codebook detailing validity criteria for the evaluation of CP interview data, and procedures for applying that codebook to rate item performance. Five raters participated in two training sessions on the application of this system; raters then rated the validity of self-report items with CP interview data collected from children. The resulting ratings were analyzed for inter-rater reliability, to assess the utility of this rating system to rate child self-report item validity.

This study employed an intraclass correlation coefficient formula—the ICC (2, 5)—to estimate reliability among the five raters. The reliability coefficient for all ratings exceeded the 0.8 benchmark, indicating a finding of substantial inter-rater reliability. That finding indicates a pattern of agreement between raters and supports the generalizability of these findings to other, similar raters utilizing such a rating system. The level of rater agreement obtained in this study is likely due to several factors: (a) the utility of rating system to rate the validity of item performance, (b) the efficacy of the rater training, and (c) the ratability of the CP interview data with respect to item performance.

The substantial level of inter-rater agreement found strengthens the utility of the ratings in the scale-development process. The ultimate goal of a CP rating system is the quantification of item performance. Those quantifications are then used to determine if an item is performing in a valid manner, and if not, whether to modify or discard the item. Therefore, there are qualitative differences between levels in the item performance scale used in this study. When raters achieve exact agreement on ratings, a clearer and more confident conclusion about the performance of the item can be drawn.

Although the reliability coefficients are well within the substantial range both for consistency and for agreement, variance can be seen in the reliability coefficients among the

five scales. For example, Scale 20 was above 0.90, which is in the upper half of the substantial range, while the coefficients for Scales 7, 17, and 18 are in the upper half of the acceptable range. The only coefficient of inter-rater reliability close to the acceptable benchmark of 0.60 was the coefficient for Scale 2, at 0.63.

This variation between the reliability coefficients seems to be partially the result of variation in the cognitive demands of items within the five scales. For instance, Scale 2 data account for four of the seven validity ratings receiving rater confidence ratings below 3.0, and three out of five of the items in Scale 2 include conditional statements. A conditional statement is a clause that designates a specific time, place, or person for the context of the item. From an information-processing perspective, items with conditional statements are more cognitively demanding for children to process (Woolley et al., 2004).

The cognitive demand of an item with a conditional statement increases the likelihood of approaching a child's information-processing capacity. The increased cognitive demands of such items also appear to introduce ambivalent or contradictory indicators of validity into the CP data. During rater training, ratings disagreements about CP data items with conditional statements were vigorously debated. It appears that the increased cognitive demand of this type of item also increases the demands on raters in the process of evaluating the validity of a child's CP interview responses. As was discussed earlier, less reliable and accurate ratings result when data are rated in the middle of the rating scale, have ambivalent or contradictory indicators, are hard to rate, or are rated with low confidence (Fox et al., 1995; Kavanaugh, 1989).

Thus, instrument developers should exercise caution in the use of conditional statements when writing self-report items for children. However, some concepts cannot be measured without including conditional statements. Consequently, it is especially important to CP those items, and to carefully consider item length and use of vocabulary, in order to minimize the already significant information-processing demands on respondents. In addition, the order of the conditional statement and the core question impacts item performance. In the third round of CP with these five scales, all conditional statements were intentionally placed up front. Through three rounds of CP, we have observed that children experience fewer cognitive processing problems when conditional statements are placed at the beginning of items.

The high confidence reported by the raters in this study permits interpretations that supplement the substantial inter-rater reliability findings. The high confidence ratings indicate that the materials that make up the rating system—the codebook, the rater guidelines, and the procedures—were clear and effective. The high level of confidence also indicates the rater training was effective in preparing the raters to accurately and consistently apply the rating system.

The core of the rating system presented here was the codebook. Writing a codebook detailing validity performance criteria for self-report items requires scale developers to engage in two valuable activities. First, the developers of an instrument must explicitly and operationally define what each item is intended to measure. Second, instrument developers must anticipate how the intended respondent population might interpret and respond to those items.

The core elements of the codebook include (a) a definition of the concept an item targets, (b) a description of what constitutes a coherent explanation of memory appropriate to the item, and (c) a set of parameters to assess the congruence of the answer choice. A codebook appears to be a valuable endeavor in the overall scale-development process because it

requires researchers to articulate, agree upon, and operationalize the nature and intent of each item.

Looking beyond the CP methodology, a preliminary version of the codebook could be included with materials provided to expert consultants during the development of items. At a minimum, clear statements of the concept each item targets would help consultants evaluate the proposed items more rigorously. In addition, preliminary and brief statements about the three validity rating criteria might further enhance a consultant's ability to evaluate the item.

With respect to the application of the CP methodology in the development of self-report instruments, the current research demonstrates the utility of a codebook and of systematic CP analysis procedures. Those procedures included the use of five trained raters. However, the results obtained with the described CP rating system indicate that three raters may be sufficient. There are 10 possible combinations of three raters, out of the five raters included in the current study. Applying the ICC (2, 5) formula to all 10 combinations of three raters, with all data, resulted in reliability coefficients ranging from a low of 0.72 to a high of 0.84. These coefficients were all well above the acceptable benchmark.

The current research furthers the development of the CP methodology as a promising tool for advancing the validity of self-report instruments. Applying the systematic data-rating procedures reported here should result in more reliable validity findings and more replicable CP results. Such reliable and replicable CP validity findings can pave the way toward empirical demonstrations of the utility of the CP methodology in advancing the validity of self-report items. Establishing an empirical foundation for the CP methodology would support its application in the development of any self-report instrument in which the intended respondents may interpret instrument items differently than instrument developers intend. Such potential instrument respondents include many populations who are participants in social science research or who are consumers of prevention and intervention programs.

Acknowledgments

The research reported in this article was previously reported in August 2003, in the first author's dissertation, as partial completion of the degree requirements for a doctorate of philosophy in social work from the University of North Carolina at Chapel Hill. This research was conducted as part of the development of the Elementary School Success Profile (ESSP). The ESSP was developed in collaboration with Flying Bridge Technologies, with funding from the National Institutes of Health (NIH) and the National Institute on Drug Abuse (NIDA), grant numbers 1 R42 DA13865-01, 3 R41 DA13865-01S1, and 2 R42DA013865-02. The research reported here was supported by a Junior Faculty Development Award to the third author, presented by the University of North Carolina at Chapel Hill. Findings, opinions, and recommendations expressed in this manuscript are those of the authors and not necessarily those of Flying Bridge Technologies, NIH, or NIDA.

References

- Bowen, NK.; Bowen, GL.; Woolley, ME. Constructing and validating assessment tools for school-based practitioners: The Elementary School Success Profile. In: Roberts, AR.; Yeager, KY., editors. Evidence-based practice manual: Research and outcome measures in health and human services. New York: Oxford University Press; 2004. p. 509-517.
- Cronbach, LJ. Essentials of psychological testing. 5th ed.. New York: Harper Collins; 1990.
- DeMaio, TJ.; Rothgeb, JM. Cognitive interviewing techniques: In the lab and in the field. In: Schwartz, N.; Sudman, S., editors. Answering questions: Methodology for cognitive and communicative processes in survey research. San Francisco, CA: Jossey-Bass; 1996. p. 177-196.
- DeVellis RF. Scale development: Theory and applications. 2003
- Dunn, G. The design and analysis of reliability studies. New York: Wiley & Sons; 1989.

- Dyrborg J, Warborg-Larsen F, Nielsen S, Byman J, Buhl-Nielsen B, Gautre-Delay F. The Children's Global Assessment Scale (CGAS) and Global Assessment of Psychosocial Disability (GAPD) in clinical practice: Substance and reliability as judged by intraclass correlations. *European Child and Adolescent Psychiatry*. 2000; 9:195–201. [PubMed: 11095042]
- Foddy W. An empirical evaluation of in-depth probes used to pretest survey questions. *Sociological Method and Research*. 1998; 27:103–133.
- Forsyth, BH.; Lessler, JT. Cognitive laboratory methods: A taxonomy. In: Biemer, PP.; Groves, RM.; Lyberg, LE.; Mathiowietz, NA.; Sudman, S., editors. *Measurement errors in surveys*. New York: Wiley & Sons; 1991. p. 393–418.
- Fox S, Bizman A, Hoffman M, Oren L. The impact of variability in candidate profiles on rater confidence and judgments regarding stability and job suitability. *Journal of Occupational and Organizational Psychology*. 1995; 68:13–23.
- Jabine, TB.; Straf, ML.; Tanur, JM.; Tourangeau, R. *Cognitive aspects of survey methodology: Building a bridge between disciplines*. Washington DC: National Academy Press; 1984.
- Jobe JB, Mingay DJ. Cognitive research improves questionnaires. *American Journal of Public Health*. 1989; 79(8):1053–1055. [PubMed: 2751028]
- Kavanaugh, MJ. *Performance rating accuracy improvement through changes in individual and system characteristics*. San Antonio, TX: State University of New York at Albany, Research Foundation; Sponsored by the Air Force Human Resources Lab; 1989.
- McDonald, RP. *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum; 1999.
- McKay RB, de la Puente M. Cognitive testing of racial and ethnic questions for the CPS supplement. *Monthly Labor Review*. 1996:8–11.
- Natsuti JP, Pecora PJ. Risk assessment scales in child protection: A test of the internal consistency and inter-rater reliability of one statewide system. *Child Welfare*. 1993; 29:28–33.
- Rebok G, Riley A, Forrest C, Starfield B, Green B, Robertson J, et al. Elementary school-aged children's reports of their health: A cognitive interviewing study. *Quality of Life Research*. 2001; 10:59–70. [PubMed: 11508476]
- Shrout PE. Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*. 1998; 7:301–317. [PubMed: 9803527]
- Shrout PE, Fliess JL. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*. 1979; 86:420–428. [PubMed: 18839484]
- Sirken, MG.; Herrmann, DJ.; Schecter, S.; Schwartz, N.; Tanur, JM.; Tourangeau, R., editors. *Cognition and survey research*. New York: Wiley-Interscience; 1999.
- SPSS. (Version 10.0.5). Chicago: SPSS Inc.; 2000.
- Weigle SC. Using FACETS to model rater training effects. *Language and Testing*. 1998; 15:263–287.
- Willis, GB.; DeMaio, TJ.; Harris-Kojetin, B. Is the bandwagon headed for the Promised Land? Evaluating the validity of cognitive interviewing techniques. In: Sirken, MG.; Herrmann, DJ.; Schecter, S.; Schwartz, N., editors. *Cognition and survey research*. New York: Wiley-Interscience; 1999. p. 133–154.
- Woehr DJ, Huffcutt AI. Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*. 1994; 67:189–205.
- Woolley ME, Bowen NK, Bowen GL. Cognitive pretesting and the developmental validity of child self-report instruments: Theory and applications. *Research on Social Work Practice*. 2004; 14:191–200.

[illegible]

Figure 1.
Codebook Sample Page

Table 1

Descriptive Statistics of Item Ratings

Rater	N	Range	Mean	SD	Skewness
Rater 1	169	0 to 3	2.66	0.84	-2.35
Rater 2	168	0 to 3	2.44	0.95	-1.51
Rater 3	167	0 to 3	2.54	0.97	-.188
Rater 4	169	0 to 3	2.51	1.02	-1.80
Rater 5	167	0 to 3	2.51	0.84	-1.82
All Raters	840	0 to 3	2.53	0.95	-1.83

Note. The item performance scale ranges from 0, total failure of the item, to 3, valid performance, which meant the subject (a) comprehended the concept of the item, (b) recalled and relayed coherent information from memory, and (c) chose an answer option congruent with the information given.

Table 2

Inter-Rater Reliability Coefficients

Data	Number of Items	ICC (2, 5) *	95% CI **
All Items	29	0.85	0.81 to 0.89
Scale 2	5	0.63	0.35 to 0.82
Scale 7	6	0.77	0.63 to 0.87
Scale 17	7	0.75	0.56 to 0.88
Scale 18	4	0.75	0.56 to 0.88
Scale 20	7	0.91	0.86 to 0.95

Note. All five raters were included in all coefficient calculations.

* The ICC (2, 5) model is a measure of rater agreement; error variance due to calibration differences between raters is included in the error variance.

** CI stands for *confidence interval*, which here is a range with a 95% chance of including the true reliability coefficient.

Table 3

Rater Confidence Ratings

Rater	N	Range	Mean	SD
Rater 1	169	2 to 4	3.87	0.35
Rater 2	166	3 to 4	3.14	0.35
Rater 3	169	3 to 4	0.48	0.50
Rater 4	167	2 to 4	3.49	0.51
Rater 5	167	2 to 4	3.54	0.55

Note. The confidence rating scale has a 4-point range from 1 (*low confidence*) to 4 (*high confidence*).