CONSIDERATIONS FOR IMPLEMENTING A CHILD ASSESSMENT IN A FIELD SURVEY RESEARCH PROJECT

Susan Sprachman, Welmoet van Kammen, and Margo Salem, Mathematica Policy Research, Inc.
Susan Sprachman, MPR, PO Box 2393, Princeton, NJ 08543-2393

Key Words: Measurement, Methodology, Cognitive Development, Bayley Scales of Infant Development-Mental Scale (BSID-II)

Using clinical assessments to measure the developmental functioning of infants and toddlers in large surveys has become more common in recent years. Fourteen years ago, the National Longitudinal Survey of Youth broke (NLSY79) new ground by adding developmental testing of the children to their data collection. Since then, in measuring program impact, many studies of women in poverty, women on welfare, and pregnant teenagers have included direct assessments of children in their protocols. While some of these studies have included school-aged children, many have included assessments of preschoolers and even infants, age groups that present special administrative challenges.

Including developmental assessments of children, especially very young ones, in large surveys creates some unique challenges. Because not all clinical assessments can be standardized to fit the needs of a field data collection effort, the instrument must be chosen carefully. Researchers need to consider such practical issues as interviewers' level of education, how best to train field interviewers to administer clinical assessments, and how to establish reliability and maintain quality control. While this paper focuses on the implementation of the Bayley Scales of Infant Development-Mental Scale (BSID-II) better known as "the Bayley" in a national evaluation of Early Head Start (EHS) programs, we believe our experiences can be generalized to help inform others who are implementing clinical assessments as part of their research survey projects.

The National Evaluation of Early Head Start is a five-year evaluation sponsored by the U.S. Department of Health and Human Services, Administration for Children, Youth and Families. In 16 different locations, the Bayley was administered to 3,000 children at three points: when they were 14, 24, and 36 months old. The Bayley tests developmental abilities by giving an infant or toddler a series of tasks, each to be administered in a precise manner. The Bayley was administered in the home as part of larger assessment that included an interview with the primary caretaker of the child, as well as videotaped interaction between parent and child. For administering the Bayley, the study did not require sites to hire specialists with professional training in individual assessments, but instead allowed the use of data collectors with a variety of educational backgrounds.

Initial Questions to Ask When Selecting an Assessment

Often survey professionals can choose between several clinical assessments that measure essentially the same domain, though in slightly different ways. Besides investigating pilot-testing procedures, bias testing of scales and items, and standardization of scales, they should determine whether the instrument has been included in other studies that required the administration of the assessment in a similar setting with a similar level of interviewing staff. The most comprehensive or advanced assessment may not always be the best one for a large field survey. When considering an instrument, the survey professional may choose to contact its author or publisher to explore the feasibility of including it in the planned survey. Investigators may want to ask if they will be allowed to change testing forms, modify training, and adjust the administration of the instrument to the particular needs of the survey.

Other questions may focus on various versions of the instrument: Has a new version of the test been issued recently? Can the results of studies using an earlier version be compared with studies using the new version? Are the authors planning to revise the instrument, and, if so, when will they publish the revised version? If investigators are planning a longitudinal study, the publication of a new version of the instrument in the middle of the study may force the researchers to adopt the revised version into a later data wave and make comparisons with data collected with an older version problematic or impossible.

Survey professionals may want to ask whether a Spanish translation of the instrument is available, whether the cost of forms and training manuals can be reduced when ordering materials in large quantities, and whether a help desk or hot line exists to answer questions in a timely manner. One does not want to be halfway into a study to discover that nobody can be found to answer even the most basic questions about the assessment.

Also important to investigate are the quality of the administration and of the scoring manual. Are the instructions for each item of the test specific and detailed

enough to enable standardization in administration? Do scoring instructions include guidelines on how to deal with assessment items that are refused, accidentally omitted, or not included because of unusual circumstances, such as when administering items to children with physical and mental challenges?

"Translating" the Assessment to Be Used in the Field Survey

Clinical assessments are intended to be used by professionals and to be conducted in laboratory or office settings. In a research project where data across a range of interviewers and sites will be compared, it is critical that the test items be administered in a standardized manner. While a clinician might take some liberties to bring out the best performance in a child, such license cannot be granted to field interviewers. Because the structure of the assessments, the testing instructions, and the format of the testing forms are often not well suited for field interviewers, the survey professional must ensure the proper "translation" of the clinical version of the assessment for successful use in the survey environment.

The term translation, usually applied to converting something from one language to another, refers here to changing the format and language of an assessment designed for a clinician to something a lay interviewer can understand. A good translator considers not just the words of a document or story, but the audience. A good translator takes into account the cultural interpretation of a phrase, and does not just translate individual words and phrases verbatim. The task facing the survey professional is thus analogous to that of a good translator.

In our translation of the Bayley, we were faced with a 375-page manual that included instructions for administering specific items,[1] as well as for determining when to stop the administration and when to administer earlier items. The test is scored on a sheet (the Mental Scale Record Form) that lists the items in order of difficulty, but not necessarily the order they are administered, which is based on grouping like items or items using the same materials and is contained in an appendix in the manual. We worked closely with the test developers at the Psychological Corporation to spell out every nuance of how each item should be administered. To make the Bayley more interviewer-friendly, we reformatted the materials and made the administration more structured, enforcing a standardized approach. These changes are described next.

Reformatting. For every item, we created a one-page card that contained the list of materials used in the item, the instructions given in the manual, and additional clarifications that our assessment consultants suggested. The cards, in the mandated order of administration, were put on a flip chart that contained all the items that we wanted administered to the children in our study. We made a supplemental flip chart for children who required the administration of items below our standard starting items and inserted a checkpoint into the main flip chart before items that would be administered only to older or higher-scoring children. To facilitate the scoring, we placed scoring boxes on each page so the interviewers could flag the appropriate score and afterwards transfer the information onto the scoring form.

Structuring for Standardization. While the Bayley items are listed on the score sheet in order of difficulty, we wanted to simplify the test administrator's task by having items using similar materials presented together, as suggested in the manual. We bound the flip chart together in this order and included a cross-reference of flip chart pages to item numbers.

We wanted to ask all children in the study a core set of items. Thus, irrespective of the child's age, the starting point for administering the test was the same. As a result, the first items administered were usually intended for children younger than the one tested and served as a warm-up, allowing the interviewer more time to establish rapport with the child.

Qualifications of Test Administrators

Hiring field interviewers able to negotiate cooperation from caretakers, conduct standardized interviews, and also administer child assessments can be a challenge. On the one hand, we ask for interviewers who are able to persuade respondents to partake in an interview; on the other, we want interviewers able to gain the trust of very young children and get them to cooperate in the assessment.

Because the Bayley is so widely used, our first inclination was to start looking for experienced Bayley administrators. However, such persons are hard to find and costly to employ, and we also realized that they would have difficulty adjusting to the study-specific structure of administering the test and following our standardized protocol of administering each scale item. In other words, they would have to unlearn old habits and techniques that may have worked well for them in testing children in their clinical practice. Also, we questioned how these assessors would respond to testing children in their home environment, where interruptions from other household members would be more the norm than the exception. In many instances, we discovered that experienced field interviewers with good people skills and the ability to follow instructions and testing rules were just as trainable in child development

programs as graduate students.

Some sites in our study opted to split the assessment between different interviewers, with one group specializing in the parent interviews and the other being trained in the Bayley and the videotaped parent-child interaction. While this strategy does enable some interviewers to focus entirely on the child assessment and may make training and maintaining quality control easier, it will be more difficult to find times that the participants and two interviewers instead of one are available to complete the assessments.

Training Considerations

Among the considerations in designing a comprehensive training program for assessments are, Who will do the training? What kind of demonstrations do you need to do? What kinds of training videotapes can you afford to produce? Do you need to arrange for hands-on training with real children? We consulted with a team of postdoctoral researchers from New York University who had extensive experience administering the Bayley to young children in research settings. The team worked with us in designing and implementing the training and the subsequent interviewer certification.

While trainees can do paired practice interviews that replicate an actual interview, it is virtually impossible to imagine what it is like to administer a task to a one- or two-year-old. Rehearsing with another adult is clearly acting, not true practice. It is useful for developing technique, but not for experiencing the actual task. In designing the Bayley training, we wanted the interviewers to have experience with a broad range of children--uninterested children, children with their own ideas of what to do with the material, children unable to sit still, and children with intruding siblings. We designed the training to include a range of experiences: observing the trainers demonstrate tasks, discussing the tasks, practicing in pairs, viewing a range of difficult-situation administrations, and, finally, practicing with real babies and receiving feedback.

Videotapes. Our NYU consultants videotaped their administration of the Bayley to 10 children. From these tapes we selected examples both of easy administrations and of more difficult ones. Our surprising difficulty in identifying what we considered to be a "perfect" administration underscored the complexity of seemingly straightforward tasks. Since we had examples of imperfect administration, we built a critique of the administration into the training activities. While videotapes can be expensive to develop, we feel that the training would have been incomplete without the opportunity to experience, in advance, some of the difficulties that the interviewers would encounter.

Conducting the Training

In sum, the actual training for the items was a mix of watching a good administration, discussing it, and practicing with the materials. Items were presented in groups so that like items were presented and practiced together and so that a training rhythm was established. We often, but not always, included watching videotapes of items that were incorrectly administered so that interviewers could develop a critical eye regarding their own administration.

We brought real babies into the training, and pairs of interviewers practiced administering the tasks with them. Finding enough babies, scheduling their visits around nap or feeding times, and fairly apportioning them to maximize practice without overtiring them can be exhausting. However, we know no substitute for this kind of hands-on practice. The most confident interviewers in training were sometimes the most insecure when confronted with a real baby, and nervous interviewers gained confidence from their success in administering items and engaging the baby. Doing their first hands-on practice in the safe haven of training helped most interviewers gain confidence in their ability to administer the Bayley tasks.

We were never able to schedule enough babies so that each interviewer could do a full practice. Often two or three interviewers took turns administering groups of items. Some of the babies were real troupers, willing to be confronted by various strangers fumbling with materials; other babies crumbled within minutes, making interviewers deal with the very real situation of not getting flustered while an embarrassed mother tried to calm her baby.

The practices were videotaped and were reviewed by the training team so that the interviewers could get feedback on their main errors right away. The group was also instructed on tasks and techniques that other interviewers had found to be problematic. Thus, by the time interviewers left training, they were aware of their most serious errors and could concentrate on refining their administration.

Certification Process

In-person training and immediate review of interviewer practice are the first steps in a lengthy certification process. No one can learn to administer the Bayley with just two days of classroom training and a half hour with a real baby. After training, interviewers were required to practice the Bayley and videotape themselves administering it to age-appropriate children. The NYU team developed an evaluation form for use both by the interviewers, to assess their own Bayley

administration, and by the reviewers. Each aspect of an item's administration was assigned one or two points based on difficulty. Individual tasks might be worth anywhere from four to seven points, depending on complexity. Interviewers scored their own administration of each item, and a member of the NYU team then reviewed and scored the tape. When we embarked on the certification, we did not require the self-evaluation, which caused many reviewer hours to be wasted on tapes that were clearly unacceptable. Once we required the self-critique, interviewers had to review their work and consider whether it was close to certifiable. The level of the tapes we received after imposing this requirement improved considerably, and the number of tapes that had to be reviewed before an interviewer was certified decreased.

Exhibit 1 contains a sample item from the evaluation form.

In this example, the interviewer thought that she had scored the item correctly, but the reviewer disagreed and felt the child should not have received credit for the item. Each interviewer was required to score 85 percent or above on two tapes to be certified to administer the Bayley to children in the sample.

Lessons Learned and Cautions

Most interviewers enjoyed the training, completed all the certification requirements, maintained reliability in administering the assessment during the study, and found the administration of the instrument relatively easy after some practice and with a little experience. While developing the training tapes was a long, laborious task, they became extremely valuable in the later years of the study when we switched to on-site training and new interviewers used the tapes to master the correct administration of each Bayley item.

We revised our recertification strategies that required each interviewer to videotape a Bayley administration after completing a certain number of completed assessments. We realized that some data collectors administered the Bayley so infrequently that more regular certifications were necessary for effective monitoring of the quality of their assessment skills. We also realized that the lag time between assessments being conducted in the field and those being received by the data collection office made it difficult to determine if the correct assessment was being videotaped for review. Midway through the study, we switched to periodic videotaping of a Bayley administration for all interviewers conducting Bayley assessments, irrespective of how frequently they had done the assessment in the previous period.

We trained assessors on item sets appropriate for the ages of the children in study. We did not anticipate that some children, because of scheduling problems, would eventually be tested outside this age window. These assessments included item sets not covered in training and certification and, because the interviewers were not qualified to administer them, created some data loss. Training interviewers in more item sets than necessary is expensive, and adhering to the assessment window, especially when studying early child development, is important. However, data collections never go exactly as planned, and including a large set of items to cover a wider age range may have prevented some data loss.

Although we carefully studied the scoring procedures as specified in the manual, we did not determine all scoring rules until we had completed almost half the assessments. If we had determined at the beginning of the study how we were going to deal with missing data, refused items and other irregularities, we probably would have been able to sharpen some of the administration rules and would have asked interviewers to fill out forms differently.

For obvious ethical reasons, we had an obligation to report to parents when the administration indicated a possible severe developmental delay in a child. Although this happened rarely, it is important to develop reporting

EXHIBIT 1

	Points	Interviewer	Reviewer
6. Rings Bell Purposely			
Rang bell in front of child	2	2	2
Put bell on table	1	1	1
Readministered correctly only if child did not pick up bell	2	2	2
Scored correctly	1	1	0
Total Item Score	**6**	**6**	**5**

procedures and a script to be used for such occasions. Interviewers need to understand that the assessments are conducted for research purposes and that these tests by themselves have no clinical value. Study directors should explain to parents that the results may indicate a potential developmental problem that would require evaluation by a developmental specialist.

Acknowledgements

The authors would like to acknowledge the help of James S. Gyurke, the BSID-II Project Director at the Psychological Corporation, for responding, no matter where he was, to our calls for help in interpreting instructions for Bayley items. We would also like to acknowledge our colleagues at New York University, Amy Damast, Emily Doolittle, Tiffany Miller, Dayana Jimenez, and Martina Albright for the Bayley expertise they brought developing of the training materials, to conducting training, and to certifying over a hundred interviewers. Finally, we would like to thank the members of the Early Head Start Research Consortium; our project officers, Helen Raikes and Louisa Tarullo; and John Love, the Project Director.

[1] Bayley, Nancy. *Bayley Scale of Infant Development,* Second Edition: Manual. San Antonio: The Psychological Corporation and Harcourt Brace and Company, 1993.