Implementing Large-Scale Studies of Children Using Clinical Assessments

Elizabeth T. Spier, Susan Sprachman, and Cassandra Rowand[1]

Research on children comes in many guises. A researcher can gain some understanding of how child development is influenced by the environment—such as parental depression or experiences in day care—by collecting data from caregivers. Substantially more information may be gained by administering developmental assessments directly to the child. However, adding an assessment to a research project can be difficult and costly when the child is an infant, toddler, or preschooler. Only a small number of tools are available for the direct assessment of cognitive and motor development in infants and toddlers. As a result, researchers often limit their choices of outcome measures or restrict themselves to studies using small samples because they are concerned that they will not be able to reliably administer complex assessments to large numbers of children. While this is a valid concern, the authors of this paper encourage researchers to think creatively in considering how to train a broad range of field data collectors to reliably administer such complex developmental assessments for large-scale studies so that we can enhance our understanding of factors influencing children's early development.

This paper highlights three case studies in adapting a complex child development scale, the Bayley Scales of Infant Development-II (BSID-II), for large-scale research projects. We first present the experience of developing a training plan for the Early Head Start Research and Evaluation Project, sponsored by the U.S. Department of Health and Human Services, Administration for Children and Families (see Administration for Children and Families 2002). Early Head Start is a two-generation program designed to provide high-quality child

[1]Elizabeth Spier is an independent international education consultant currently residing in Amsterdam (elizspier@aol.com).; Susan Sprachman, the corresponding author, is senior survey researcher at Mathematica Policy Research, Inc., Princeton, NJ (ssprachman@mathematica-mpr.com); Cassandra Rowand is a survey

and family development services to low-income pregnant women and families with infants and toddlers, and now operates in more than 700 communities around the country. Mathematica Policy Research, Inc. (MPR) conducted the national evaluation of this new initiative, assessing the impacts of the program on a wide range of child outcomes, parenting and the home environment, and family self-sufficiency on a sample of approximately 3,000 families.

The second case study shows how the Bayley Scales of Infant Development were converted into a valid and reliable "short form" for administration to 13,500 children for the Early Childhood Longitudinal Study—Birth Cohort (ECLS-B), sponsored by the U.S. Department of Education, National Center for Education Statistics (NCES). The ECLS-B, conducted by Westat, Inc., will follow the development of a nationally representative sample of approximately 13,500 children born in the year 2001 from birth through first grade.

Finally, we present the challenges of "exporting" the BSID-II to evaluate the impact of swaddling on development among 1,300 children in Mongolia—a country in which such assessment tools have been little used to date. The Investigation of the Effect of Swaddling on Lower Respiratory Infection in Mongolia Infants project was funded by the Wellcome Trust and the Canada Fund.

The BSID-II consists of a Mental Scale and a Motor Scale for children aged 1 through 42 months; either one or both scales may be used. Administering the BSID-II involves the complex and precise manipulation of a complicated flow of testing materials and instructions, rules for which items to administer based on the child's age and performance, and judgment calls about whether credit should be given or not. Clearly, the decision to use an instrument like the BSID-II with a staff including individuals who are inexperienced with the measure

---

researcher at Mathematica (crowand@mathematica-mpr.com). This paper was prepared for the Children in the Mediterranean Conference, Genoa, January 7-9, 2004.

should not be made without carefully considering the ability of the project staff to develop a comprehensive administration and training plan. We feel that presenting these three examples of using the BSID-II will encourage more researchers to think creatively about how they might incorporate complex assessments into large-scale research projects.

Adapting the BSID-II for the Early Head Start Research and Evaluation Project

When a complex measure is to be administered by a field staff with varying backgrounds, it is crucial that careful attention be given to making the assessor materials as user-friendly as possible. For example, the field staff for the Early Head Start study was drawn from a range of individuals and we could not assume that any staff had prior experience with developmental assessments. We will discuss below how, for the Early Head Start study, staff from MPR and New York University (NYU) carefully designed the instructions we would provide the assessors, reformatted the instructions and scoring sheet, made the basal and ceiling calculations less prone to error, provided direct training to the staff, and ensured that they were administering the test reliably.

"Translating" the Assessment for the Field Survey: For such assessment tools as the BSID-II, the test structure and instructions, and the format of the scoring forms, often are not well suited for use by field interviewers. Once we made the decision to use the Mental Scale from the BSID-II, our first task was to "translate" 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 merely translate individual words and phrases word for word. Thus, the task facing the survey professional is analogous to that of a good translator.

In our translation of the BSID-II, we were faced with a 375-page manual that included instructions for administering test items, as well as for determining which items to administer based on the child's age and performance. The child's performance is scored on BSID-II "Record Forms" that list the items in order of difficulty, but not necessarily the order in which they are administered. The order of item administration is based on grouping like items or items using the same materials and is contained in an appendix in the manual. To make the BSID-II Mental Scale more interviewer-friendly, we reformatted the materials and made the administration more structured, enforcing a standardized approach. We wanted to ask all children in the study a core set of items. Thus, regardless of the child's age (the range was 13 to 16 months), the starting point for administering the test was the same. In most instances, this resulted in first administering items intended for children younger than the ones tested, as a warm-up, allowing the interviewer more time to establish rapport with the child.

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 suggested by our assessment consultants. While the BSID-II 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. The cards, in the mandated order of administration, were put in a flip chart that contained all the items we wanted administered to the children in our study. We bound the flip chart together in this order and included a cross-reference of flip chart pages to item numbers. We made a supplemental flip chart for children who required the administration of items below our standard starting items and inserted a checkpoint in the main flip chart before items that would be administered only to

older or higher-scoring children. The cards contained the basic instructions from the manual, as well as administration hints. We placed scoring boxes on each page so the interviewers could easily flag the appropriate score and afterwards transfer the information onto the BSID-II Mental Scale Record Form.

Training the Research Staff: Among the considerations in designing a comprehensive training program for assessments are who will do the training, what kind of demonstrations need to be done, what kinds of training videotapes can be affordably produced, and whether arrangements need to be made for hands-on training with children. The MPR-NYU team worked together in designing and implementing the training and the subsequent interviewer certification.

While trainees can do practice interviews in dyads that replicate administering a questionnaire or a test to an adult, 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 useful for developing technique, but not for experiencing the actual task. In designing the BSID-II training, we wanted the interviewers to have experience with a broad variety of children—uninterested children, children with their own ideas of what to do with the materials, restless children, 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 babies and receiving feedback.

Our consultants videotaped their administration of the BSID-II, andfrom these tapes we selected examples of both easy administrations and 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 an imperfect

administration, we built a critique of the administration into the training activities. 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 a rhythm could be established. We also included watching videotapes of items that were incorrectly administered so that interviewers could develop a critical eye regarding their own administration.

We brought 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 actually confronted with a baby, and nervous interviewers gained confidence from their success in administering items and engaging the baby in the safe haven of training. We were never able to schedule enough babies so that each interviewer could administer a full practice test. 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.

Ensuring Proper Administration*:* 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 full BSID-II for 14-month-olds with just two days of classroom training and a half-hour with a baby. After training, interviewers were required to practice the BSID-II Mental Scale and videotape themselves administering it to age-appropriate children. When we embarked on the certification, we did not require self-evaluation, which caused many reviewer hours to be wasted on tapes that were clearly unacceptable. The NYU team thus developed a critique form that could be used first by interviewers to review their administration and consider whether it was close to certifiable, then checked by the reviewers. 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.

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. Each interviewer was required to correctly score 85 percent or above on two tapes in order to be certified to administer the BSID-II Mental Scale to children in the sample. If they scored between 70 and 85 percent on the second tape, interviewers received a conditional pass allowing them to administer the BSID-II with research children but requiring them to videotape additional BSID-II administrations for certification purposes until they had reached the required standard of administration.

Developing the Bayley "Short Form" for the Early Childhood Longitudinal Study

The Early Childhood Longitudinal Study—Birth Cohort (ECLS-B) focuses on how child and family characteristics, early health care, and in-home and out-of-home child care history influence children's first experiences with the demands of formal schooling. Measures of children's early physical, cognitive, social, and emotional development were taken during home visits when the children were approximately 9 months of age; subsequent

data collections are planned for when the children reach 24 and 48 months of age, and upon entry into kindergarten and first grade.

The successful adaptation of the BSID-II for the Early Head Start study enabled this project to propose its use in the ECLS-B. The ECLS-B differed from the Early Head Start study, in that the BSID-II Motor Scale was added, and that testing would begin when children were 9 months of age rather than 14 months. Thus, the materials developed for the Early Head Start study had to be modified for use in the ECLS-B. We designed assessment booklets and scoring sheets for the ECLS-B based on the format described above for the Early Head Start Study but covering a range of items appropriate for 9-month-olds. The Motor Scale items followed the Mental Items in the booklet.

The first ECLS-B field test demonstrated that a full administration of the BSID-II, even when streamlined and condensed through the use of flip charts, required at least 40 minutes for the Mental Scale, plus time for the Motor Scale. Upon completion of both the Metal and Motor scales "to ceiling," children had often reached their limit and were unable to continue with the other activities required for the study. In addition, the protocol included 10 other tasks so that the entire home visit required in excess of 90 minutes to complete. Such a lengthy home visit was beyond the scope of the project.

The BSID-II is considered the gold standard in standardized early developmental assessment, and NCES was unwilling to eliminate it from its protocol. In its original state, however, the BSID-II took up far too much time to be practical for inclusion in the ECLS-B study. The challenge, then, was to create a "short form" so that valid, reliable BSID-II Mental and Motor scale scores could be obtained for the children in the study, while substantially reducing the time required to complete the assessment. The goal was a

maximum administration time of 25 minutes for the Mental Scale and 15 minutes for the Motor Scale.

Adapting the Assessment for the Field Survey: The ECLS-B contractor, Westat, was charged with developing a reduced set of items that could be administered in less time and produce reliable, valid scores equivalent to the full set of BSID-II Motor and Mental scale items. In the end, the "Bayley Short Form" was developed, field tested, refined, and implemented in the nine-month data collection of the ECLS-B. Selection of the final items was a multi-step process. First, Item Response Theory (IRT) analyses identified the strongest and most predictive items using the original, nationally normed BSID-II database developed by the Psychological Corporation. Second, BSID-II experts looked at those items that were found to be most predictive, and judged both the administration time of the items and the expected difficulty in training field staff to administer them. They found that the use of fewer materials and transitions during testing could reduce administration time. And, finally, child development experts examined the reduced set of items for thoroughness in covering all essential developmental domains. Assessment booklets and scoring sheets were assembled containing the reduced set of Mental and Motor scale items that comprise the Bayley Short Form.

The Mental Scale of the 9-month Bayley Short Form consists of a core set of 15 items—compared to 45 items for the Early Head Start 14-month assessment. The Motor Scale had a core set of 10 items. New basal and ceiling rules also had to be established for the Bayley Short Form. The Bayley Short Form Mental Scale requires three or more items correct to establish a basal, and three or more items incorrect to reach a ceiling, compared with five or more items correct to reach a basal and three or more items incorrect to reach a ceiling for the full BSID-II. Two or more correct items on the Bayley Short Form Motor

Scale establish the basal, compared with four or more for the standard BSID-II, and two or more incorrect items establish the ceiling in both the Bayley Short Form and the BSID-II. As in the Early Head Start BSID-II administration, a checkpoint was established at the end of the core set of Bayley Short Form items for testers to determine whether they needed to administer additional items to reach a basal or ceiling rule for that child.

Training the Research Staff: Training field staff to administer the Bayley Short Form Mental and Motor scales was conducted over three days, with a day for practice with a child in a format similar to that used to train the testers for the Early Head Start study. A training video was used to highlight the administration of items and scoring nuances. A video clip was shown with the administration of each item, followed by correct and incorrect responses. Field staff were then given an opportunity to administer the items in dyads, with one person acting as the child. When trainees became familiar with the items, they watched a video administration of the Bayley Short Form with a nine-month-old child. This tape was paused following the administration of each item, and trainees were asked to critique the tester's administration of the item and to score the child's response for that item. A critique form similar to that designed for the Early Head Start study was designed to review testers' administrations.

Ensuring Proper Administration: Field staff were required to be certified to administer the Bayley Short Form in a manner similar to that employed in the Early Head Start study. They worked in pairs with one videotaping while the other assessed and then switched off for the second session with a new child. On the last day of training, children between 8 and 10 months of age were brought in, and each trainee administered the Bayley Short Form to a child. Administrations were videotaped and critiqued by training staff, using the form mentioned above. Trainees were required to meet a certification level of 85 percent on the

Mental and Motor scales, independently. Trainees who did not reach 85 percent at training on both scales were required to videotape an assessment with a nine-month old after they attended training. To guard against assessor drift, periodic reliability checks were made throughout the data collection period. Trainees were required to meet a certification level of 85 percent.

Adapting the BSID-II for the Infant Swaddling Study in Mongolia

The BSID-II Mental and Motor scales are currently being employed to examine some of the risks and benefits to development of tight swaddling among infants in Mongolia. Although field staff were trained to use the BSID-II specifically for that study, the training was also intended to serve as a foundation for the future application of the BSID-II as a clinical tool in Mongolia. Therefore, unlike the staff for the Early Head Start and ECLS-B studies, the Mongolian trainees were all pediatric professionals rather than laypersons. There is a plan to develop BSID-II Mental and Motor Scale norms for 13-month-olds based on the large swaddling study sample, and it is hoped that Mongolian BSID-II norms for other ages can be established in the future. We used the materials, training, and certification structure from the Early Head Start study as a basis for the training in Mongolia. However, additional considerations arose in extending the BSID-II to this new cultural setting, and the BSID-II training was more rigorous and extensive in Mongolia to ensure that these issues were adequately addressed.

"Translating" the Assessment for the Field Survey: The BSID-II had to be translated into Mongolian in the literal sense, but that was only the beginning of the process to ensure optimal understanding of the test. In fact, during the two weeks of training, translation became an iterative process by which BSID-II instructions in Mongolian were refined as the testers became more familiar with the focus and nuances of each task. We found it necessary

over the course of training to set aside time in several sessions specifically for the testers to discuss the translation as a group and come to an agreement as to the best wording.

The flip cards for the BSID-II Mental Scale that had been used at the 14-month assessment for the Early Head Start project were translated into Mongolian. In addition, Mongolian staff found it helpful to incorporate pictures of the materials needed for each task onto the relevant card, to facilitate quick transitions from one task to the next. In addition, a new set of cards was made for the BSID-II Motor Scale items to be used in the research. The cards were originally made in English, in the same basic format as the Mental Scale cards, then translated into Mongolian.

Training the Research Staff: The training was conducted over 10 full days at the hospital where the research staff worked in the Mongolian capital, Ulaanbaatar. Trainees were hospital pediatric staff who had extensive experience with young children, but who were unfamiliar with this type of developmental assessment. The trainer spoke English, with translation into Mongolian by a pediatrician participating in the training. The first few days of training were devoted to explaining the purpose of the BSID-II, understanding standardized testing in general, watching videotapes of sample administrations, and learning what the BSID-II materials were and how to care for them. This last item seems minor, but it was especially important in Mongolia because if materials became lost or broken, it would be prohibitively expensive and time-consuming to order replacements, and substitutions with local materials would not be acceptable for standardized administration.

The trainees practiced administering the test to each other in pairs for two days, then had five full days of practice with children the age of study participants. A child would be brought into the training room, and one tester would administer the Mental and Motor scales while the other testers and trainer watched. The group then discussed what they had

observed in terms of completing the tasks exactly according to instructions, the overall engagement of the child, smoothness of administration, and so on. It also presented an opportunity for the group to discuss how they would have interpreted ambiguous responses on the part of the child, and allowed trainer to clarify instructions and ensure adequate understanding of each task by the group as a whole. The trainees then split in two groups and completed some additional BSID-II administrations in children's homes, in order to gain a more "real life" experience of testing where siblings were present (for example, perhaps there was no table to sit at with the child). We found that the testing became much more precise and reliable when the examiners had the opportunity to discuss why tasks were administered in a certain way. "Because the manual says so" was not accepted as a sufficient explanation in this setting. For example, the testers informed the trainer that they could not see why you would follow the BSID-II manual instructions and say, "See my tower?" to a one-year-old who had no idea what a tower was. Therefore, they would not bother to use the wording at all while administering that item, or would say it at the wrong time. In discussion with the trainer, the Mongolian staff agreed that children of that age could indeed associate this label with a sample block construction the examiner was making if uttered concurrently, and that hearing "tower" again while more blocks were placed in front of her would suggest to the child that she was to build in the same way. Once that connection was made and the Mongolian testers understood that the wording to the child was not superfluous, then they consistently administered the task exactly according to the instructions in the manual.

Ensuring Proper Administration: The certification process for the testers in Mongolia was very similar to the certification for research staff in the Early Head Start study, but more stringent, for several reasons. First, it was expected that the training would eventually form a foundation for more general use of the BSID-II as an assessment tool in Mongolia beyond the

swaddling study. Second, because the testing was being conducted in Mongolia for the first time, it was especially important to make sure that the testers would manage the overall assessment procedure in a way that was comparable to what the test designers had originally intended. This often included judgments on the part of the examiner, such as whether instructions are to be repeated or how an item should be scored based on ambiguous actions on the part of the child. And third, the BSID-II Motor Scale was being applied in this manner for the first time, so it was important to make sure that its administration flowed smoothly and that all necessary reminders for proper administration and scoring had been included in the Motor Test flip chart.

In order to be certified to administer the BSID-II to children in the research study, testers had to videotape themselves practicing on age-appropriate children and fill out the same self-evaluation form for the Mental Scale as the testers from the Early Head Start study—translated into Mongolian. An additional self-evaluation form was developed to cover the Motor Scale items in the same manner. Each tester had to submit two certification tapes, each with scores of at least 90 percent for both the Mental and Motor scales. If a tester achieved 90 percent or above on one scale, but not the other, they had to submit an additional videotape of themselves administering the scale they had not passed until they achieved the required standard.

In Conclusion

As these case studies have demonstrated, it is clearly possible to use complex developmental assessments in large-scale studies. However, the diverse issues that arose during the three studies show that there is no “one-size-fits-all, how-to” manual for doing so. Rather, careful attention must be given to how materials can be formatted for ease and efficiency of use and to minimize errors, how research staff will be trained, how testing may

need to be adapted to fit into time constraints during field assessments, and how to ensure that accurate translation, understanding, and administration can be achieved when extending testing to novel cultural contexts—all the while remaining faithful to the original test. We hope that the examples we have provided will allow others to consider how they may undertake or enhance large-scale studies of children through the use of clinical developmental assessments.

Acknowledgments

The Bayley work conducted as part of the national Early Head Start Research and Evaluation Project was funded by the Administration for Children and Families (ACF), U.S. Department of Health and Human Services under contract 105-95-1936 to Mathematica Policy Research, Princeton, NJ, and Columbia University's Center for Children and Families, Teachers College, in conjunction with the Early Head Start Research Consortium. The content of this paper does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. The authors would like to acknowledge the help of James S. Gyrke, the BSID-II Project Director at the Psychological Corporation, for help in interpreting instructions for BSID-II items. We also would like to acknowledge New York University trainers (Amy Damast, Emily Doolittle, Tiffany Miller, Dayana Jimenez, and Martina Albright) for their help in developing the training materials, conducting training, and certifying more than a hundred interviewers. Finally, we would like to thank the members of the Early Head Start Research Consortium; our project officers, Helen Raikes, Louisa Tarullo, and Rachel Chazan Cohen; John Love and Ellen Kisker, Co-Project Directors; and Margo Salem who helped design the training program and flip chart.

Many thanks to the Westat staff who worked on the Bayley Short Form—Philip Fletcher, Gary Resnick, Nick Zill, Carol Andreasson, and Susan Gilmore. We would like to thank Jerry West at NCES and the Westat Project Director Brad Edwards, Michael Skinner of Pendragwn Productions and Charles McNeill in the Westat Graphics department.

We would like to thank the Canada Fund and Dr. Semira Manaseki-Holland for enabling us to bring the BSID-II to Mongolia, and Tsogzolmaa Bayandorj and Bayasgalantai Bavuusuren for their tremendous efforts in translation, reformatting the BSID-II materials in Mongolian, and providing all the extensive practical support needed to conduct the training.

References

Administration for Children and Families. *Making a Difference in the Lives of Infants and Toddlers and Their Families: The Impacts of Early Head Start.* Washington, DC: U.S. Department of Health and Human Services, June 2002.

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