The Impact of Privacy Breaches on Survey Participation in a National Longitudinal Survey

Marilyn Seastrom[1], Chris Chapman[1], Gail Mulligan[1]
[1] National Center for Education Statistics, 1900 K Street NW, Washington, DC 20006

Abstract
A longitudinal study with five rounds of data collection planned over a 5-year period experienced data loss during field operations in the third and fourth rounds of data collection. Cases with data loss were evaluated for risk of disclosure. All cases with data loss were notified, and the level of potential risk was reported to each case. Those with a potential risk of disclosure were offered a year of credit monitoring. Cases with data loss were allowed to opt out of the data collection and to have their data removed from the computer files. The disclosure risk analysis is described; the numbers of cases impacted are reported, along with the number of cases that opted out during the third and fourth rounds of data collection; and the participation of the impacted cases compared to the rest of the sample is compared in rounds three and four of the data collection.

Key Words: Privacy, data breaches, response rates, nonresponse bias analysis

1. ECLS Background

The Early Childhood Longitudinal Study (ECLS) is sponsored by the National Center for Education Statistics (NCES) in the U.S. Department of Education's Institute of Education Sciences. The Early Childhood Longitudinal Study includes two components—a school-based nationally representative sample of children who were enrolled in kindergarten in the 1998-99 school year (ECLS-K) and a household-based sample of children who were born in the United States in 2001 (ECLS-B).

1.1 ECLS-B Survey Components
The birth study was designed to provide detailed information on children's development, health, and in- and out-of-home experiences in the years leading up to school. Unlike most NCES surveys, the birth study was conducted in the children's homes. This nationally representative 2001 cohort of children was selected from registered births in the National Center for Health Statistics (NCHS) vital statistics system. This cohort of children was followed through kindergarten entry, with data collections at ages 9 months, 2 years, 4 years (pre-kindergarten), and then at ages 5 and 6 when these children were in kindergarten.

For the fourth round of collection in the fall of 2006, kindergarten enrollment status was determined for each child. About 75 percent of the children were enrolled in elementary school, and the other 25 percent had not yet entered school. Both groups of children were included in the fourth round. In the last round of collection, data were obtained only for the 25 percent of children who were first-time kindergartners, as well as for some children who had been in kindergarten the year before but were in kindergarten again in 2007.

Every round of data collection has included a home visit comprising a parent interview (usually with the child's biological mother) and, with the parent's permission, a child assessment. Father questionnaires were administered for resident fathers in the first three rounds of collection and for nonresident fathers in the first two rounds of collection.

During the parent interviews, the parent or guardian was asked to provide socio-demographic information about their family, along with information on their attitudes, the child's home educational activities, childcare experiences, and child development and health. Most parent interviews and child assessments were conducted in English, but they were also conducted with parents who spoke other languages. Bilingual interviewers were trained to conduct the parent

interview and assessments in either English or Spanish. An interpreter was used for families who spoke languages other than English or Spanish. As a result, less than 0.1 percent of the cases were not completed due to language difficulties.

During the assessments at 9 months and 2 years, children participated in a variety of activities that were intended to assess their early cognitive, physical, and socio-emotional development. Children's cognitive and physical skills were measured through a one-on-one untimed assessment —the Bayley Short Form–Research Edition (BSF–R)—in which a trained staff member assessed each child. Each child also had weight and height measurements taken to assess physical growth. Finally, the child was videotaped interacting with the parent to assess socioemotional functioning. An assessment of each child's attachment relationship with his or her primary caregiver (usually the mother) was included in the 2-year collection only, also to measure the child's socioemotional development.

In the third round at age 4, an assessment of children's language, literacy, mathematics, and color knowledge was administered, along with a fine motor skills assessment and physical measurements. These same assessment components, with the exception of color knowledge, were used in the two kindergarten collections. These assessments included audiotaping of children retelling stories to assess their early reading skills.

In the 2-year and 4-year rounds of data collection, additional components included an interview with the child's primary nonparental caregiver and, for a subsample, an observation of the nonparental care arrangement in which the child spent the most hours. Care provider interviews were continued in the kindergarten collections, though the observation component was not. Finally, the teachers of children who were enrolled in school in the kindergarten collections were asked to complete self-administered questionnaires.

1.2 ECLS-B Response Rates by Wave

During the first wave of the study, 13,900 sampled and fielded births yielded 10,700 cases with at least a completed interview with the child's parent and 10,200 children were directly assessed. Overall, 74.1 percent of the children in the sample had parents who participated in the first wave of data collection at 9 months, and 96.7 percent of the children with participating parents were assessed.

In the second wave of the study, parent interviews were completed for 9,850 of the 10,700 children from the first wave of the data collection, for a response rate of 93.1 percent (69.0 percent of the original eligible sample). Of children whose parents participated in the 2-year parent interview, 94.2 percent were assessed (65.0 percent of the original eligible sample).

In the third wave of the study, pre-kindergarten parent interviews were completed for 8,950 of the 9,850 children who participated in the 2-year collection, for a response rate of 91.3 percent (approximately 63 percent of the original eligible sample. Of children whose parents participated in this wave, 98.3 percent participated in the pre-kindergarten assessment (approximately 62 percent of the original eligible sample).

In the Kindergarten 2006 data collection (wave 4) parent interviews were completed for 6,450 of the eligible 7,050 cases,[1] for a response rate of 91.8 percent. Of children whose parents participated in this wave, 98.6 percent were assessed. Approximately 1,800 cases were interviewed in 2007, and results are not yet available for that collection.

These aggregate response rates only tell one portion of the response rate history for this study. In ECLS-B, the home visits included the computer assisted parent interviews, videotaping of the parent and child interactions, the assessments of the children, and height and weight measurements of the child. At each age, the assessment required a considerable amount of materials and equipment that had to be to be unloaded and taken into the house and then packed up and removed from the house at the end of the session. As a result there were numerous "opportunities" for data breaches to occur.

The pre-Kindergarten data collection was conducted between the fall of 2005 and early summer of 2006. At that time, there was an increased awareness of the risks of identity theft as a result of unauthorized access to personally identifiable information. In February of 2005, OMB issued Memorandum M-05-08, calling for the designation of

[1] For budgetary reasons, the kindergarten 2006 data collection followed a reduced sample of ECLS-B participants. The sample reduction was approximately 85 percent of children who had responded at all of the prior waves (9 months, 2 years, and preschool).

senior agency officials for privacy. Then in June of 2005, OMB issued Memorandum M-05-15 that added a series of questions about privacy programs to the annual report required under the Federal Information Security Management Act of 2002 (FISMA); included were questions on the privacy officers' roles in the agency, the availability of staff training on relevant privacy laws and procedures, and on policies for web privacy and for privacy impact assessments. This was followed by the May 2006 OMB Memorandum M-06-15 on "Safeguarding Personally Identifiable Information" that stressed each agency's responsibilities to safeguard sensitive personally identifiable information. In July 2006, OMB Memorandum M-06-19 required all agencies to report all incidents of suspected or confirmed privacy breaches involving personally identifiable information[2] to the U.S. Computer Emergency Readiness Team (US-CERT) within one hour of discovering the incident.

As a result of this increased emphasis on privacy protections and the reporting of privacy breaches, there is documentation of the potential privacy breaches resulting from the third and fourth waves of the ECLS-B data collection. Because this study used National Center for Health Statistics (NCHS) Vital Statistics birth records as the sampling frame, the study was conducted under the auspices of the NCHS Ethics Review Board (ERB). This documentation is due in part to the emphasis that the NCHS ERB placed on the reporting of adverse events. The resulting documentation includes descriptions of the sources of data loss, the number of cases affected, an assessment of the level of risk associated with each loss, actions taken to notify the affected individuals, and the resulting decisions taken by those individuals concerning their participation in ECLS-B. The fact that ECLS-B is a longitudinal study provides a unique opportunity to evaluate the effect that privacy breach notifications have on the continuing participation of the affected individuals.

2. ECLS-B Pre-Kindergarten Data Collection

2.1 Field Operations and Data Breaches

Data collection for the pre-kindergarten wave of ECLS-B started in August of 2005. Field interviewers carried case folders that included information about each study family, including an address history form, a case information sheet, and parental consent forms, as well as a laptop used to conduct the interview. The first data breach occurred on November 2, when a field interviewer left a hard copy case folder for one case at a childcare center when conducting a classroom observation. The next three data breaches, all occurring in November, involved the loss of individual laptops and electronic equipment. Other sources of data loss included a misdelivered FedEx package containing hard copy case files that were being transferred from one field interviewer to another.

Upon learning of the first of these data breaches, corrective modifications were implemented in the field, and work was stopped on the affected cases pending evaluation of the risk and a determination of the appropriate actions to take. As a result of these data breaches, data collection was suspended in early December and a full audit of all study materials and processes was conducted to identify any other lost information and potential sources of loss. It was discovered that existing protocols for obtaining informed consent from study participants, for shipping confidential information, and for securing confidential information in the field were not always followed. In addition, inadequacies were identified in the protocols for shipping confidential information, for securing confidential information in the field, and for reporting potentially missing case materials. The field procedures were revised to address the identified inadequacies (primarily through increased tracking and reporting). New training materials that incorporated the revised procedures were developed, locks were provided for the laptops, and field interviewers were not released to start working again on ECLS-B until they completed retraining and certification for data security.

During this period, risk analyses were performed for the different types of privacy breaches, and in those cases where personally identifiable information was potentially exposed, disclosure analyses were performed. Following a decision to notify families where any risk was identified, the levels of risk were sorted into three categories and telephone scripts, letters informing the family of the disclosure, and documents requesting reconsent to participate in the study

[2] Personally Identifiable Information (PII) is defined by OMB as any information about an individual maintained by an agency, including, but not limited to, education, financial transactions, medical history, and criminal or employment history and information that can be used to distinguish or trace an individual's identity, such as their name, social security number, data and place of birth, mother's maiden name, biometric records, etc., including any other personal information which is linked or linkable to an individual.

were developed. Data collection for the cases not affected by data breaches resumed in late January 2006, and the notification and reconsent process started in early April.

Despite the revised procedures and retraining efforts, more data breaches occurred. There were two additional instances of computer, video camera and audio recorder thefts. The most common data breaches involved missing Case Inventory Sheets and/or Address History Forms; although there was at least one more lost Fed Ex package, and there were two instances reported that involved missing reconsent forms. As these additional breaches were reported, the cases were removed from the active caseload; risk analyses, and where necessary, disclosure analyses were performed; and the reconsent process was initiated.

2.2 Risk Analysis, Level of Risk Defined, and Disclosure Analysis

There were different types of personal information on the different forms included in the case folder, so the risk of disclosure from a data loss was evaluated for each form. Since a public-use data file was distributed for a short time for the first wave of data collection, it was assumed that the first wave public use file was available in the public domain.

2.2.1 Risk Category I

The Address History Sheet includes the respondent's name and address, but no further information. The parental consent forms, reconsent forms, and the Child Assessment Booklet (used to record children's performance on the assessments) include the respondent's name and the name of the study. Thus, the only thing someone finding any of these materials would know is that a child participated in the study. Unless that person knows something more about the specific child that would help identify that child in the data file, there is no risk of that person identifying additional information from the study responses. Losses of these materials are assumed to involve a negligible risk of further disclosure and a negligible risk of harm to the respondent.

Lost tapes of parent/child interaction and/or audiotapes included only the first name of the child and thus are not linkable to other survey information; as a result, these losses were also assumed to pose a negligible risk. The laptops that were lost or stolen each required two passwords to access the database with case responses, and field interviewers were instructed to not keep those passwords with the machine. If that protocol was not followed and the intruder was able to access the laptop, the intruder would have needed systems knowledge and the ability to use a specific survey processing set of software (Blaise) in order to access survey responses that were stored in the laptop. Combined, these factors make it unlikely that anyone could access the data in the stolen or lost laptops. Furthermore, one aspect of the revised protocols included more frequent transmission of data from completed cases, which removed them from the laptop's memory.

The 140 cases that experienced these data breaches (and none more serious) were assigned to the lowest level of risk, where the risk was assumed to be negligible.

2.2.2 Risk Categories 2 and 3

The 41 lost Case Information Sheets included the largest amount of substantiated loss of information. These hard-copy paper forms included the name, date of birth, and race and ethnicity for the child, the mother, and the resident father. In addition, the mother's and father's education level were included, along with the marital status of the mother, the respondent's preferred language, and the language used for the last interview. For the child, the sheet also included an indication of whether the child was a twin, was of very low birth weight, and tribal information for American Indian children. There are two primary reasons to be concerned about these lost information sheets.

The first is that an evaluation of the information included on the Case Information Sheets concluded that if the lost information sheets were found by the wrong person(s), the lost information provided sufficient data for identity theft to occur. All 41 cases with lost Case Information Sheets were placed at potential risk for identity theft and were thus determined to be eligible for credit monitoring (one remedy NCES developed to mitigate damage from lost PII).

The second reason for concern is that anyone finding the lost case information sheets might decide to use the information to learn more about the respondent based on information provided in the study. Because of this potential, disclosure risk analyses were conducted against the existing ECLS-B public-use file for 9-month-olds to identify cases for which the data loss was potentially disclosive. To do this, the data that were on the case information sheet were recoded to match the categories for the same variable in the released public-use file, and the two data sets were

compared. Singletons, which are the lost cases that match only their own record on the public-use file, are at the most risk of disclosure. In contrast, for the lost cases that match their own record and at least two others, the risk of disclosure is less certain. That is, the probability of a correct identification of a case decreases as the number of exact matches that the compromised case yields increases.

Nine cases were identified as singletons with the potential to have their responses to the first wave of the ECLS-B data collection identified; these cases were also at risk of identity theft. As a result, these cases were assigned to risk category 3. In contrast, the 32 cases with multiple exact matches on the public-use file were not at risk of having their responses to the first wave of the data collection identified, but were at risk of identity theft and thus were assigned to risk category 2.

2.2.3 Summary of Risk Categories

Overall, some 2 percent, or 181, of the respondents to the pre-Kindergarten wave of ECLS-B experienced a data breach. Of these 181 data breaches, just over three-quarters (77 percent) were categorized in level 1 with a negligible level of risk of further disclosure or harm. Just under one-fifth (18 percent) of the data breaches were assigned to risk level 2 because there was the possibility of the lost information being used for identity theft. The smallest group of respondents (5 percent) were assigned to risk level 3 because there was the possibility of the lost information being used for identity theft and having their responses to the first wave of data collection disclosed.

When these 181 cases are considered as a group and compared to the full ECLS-B pre-Kindergarten sample on a set of key social and demographic variables, the data show that the data breaches appear to have occurred at random. That is to say, there are no significant differences in the distribution of the data breach cases relative to the distribution of the full sample when mother's age at child's birth, family structure, the language of the parent interview, mother's education, household poverty status, and urbanicity are analyzed. Similarly, there are no significant differences in the distributions of the same set of variables between the cases without data breaches and the full sample.

2.3 Reconsent Procedures

Each affected case was given three options: withdraw from the study and have their data for the pre-Kindergarten wave of data collection removed from the data files, withdraw from the study while leaving their pre-Kindergarten data in the data file, or remain in the study. In addition, respondents in risk categories 2 and 3 were extended an offer of credit monitoring to let them know if any activity related to identity theft occurred.

The reconsent process started in early April 2006. The protocol included an initial telephone contact to inform the respondent of the event and to alert them to expect a notification letter. Notification letters were written for each of the three categories of risk, and each affected case received the appropriate notification letter along with a copy of the reconsent form to review and sign. The letters were followed by a second telephone call to ensure the letter was received and to answer any questions or discuss any concerns. Respondents in categories 2 and 3 (and those in category 1 with serious concerns) were offered home visits to further discuss the data breach.

2.4 Reconsent Results

Overall, 40 percent of the affected cases chose to continue in the study, 15 percent withdrew from further participation, but let their pre-Kindergarten data remain in the data file, and 14 percent withdrew from the study and asked that their pre-Kindergarten data be removed from the data file. For the 22 percent of the data breach cases who did not respond to the reconsent request, their nonresponse was treated as a withdrawal from the study and any existing pre-Kindergarten data were removed from the database. The remaining 9 percent of the affected cases were not located and were also treated as withdrawals. Thus, the pre-Kindergarten data for 45 percent of the affected cases were removed from the database (14 + 22 + 9 percent).

When the results of the reconsent process are examined by risk category, the data show that two-thirds of the affected cases in the highest risk category (3) were removed from the database, compared to just over one-half (56 percent) in the middle risk category (2) and four out of ten (41 percent) in the lowest risk category (1). However, because three quarters of the affected cases were in the lowest risk category, more cases were removed from the database due to low risk data breaches than due to higher risk breaches (58 versus 6 cases).

2.5 Resulting Survey Response Rates

Overall, fewer than 100 cases (82) were lost to the pre-kindergarten data collection, and the responding cases (which includes both those affected by the loss and those not affected) comprise 90 percent of the original pre-kindergarten sample. Response rates were 91 percent for the respondents not affected by data loss and 89 percent for the respondents who were affected by data loss.

As mentioned above, 15 percent of the affected cases agreed to leave their data in the pre-kindergarten wave but withdrew from further participation in the kindergarten waves of the data collection. Thus, the pre-kindergarten data breaches had a greater impact on sample loss in the subsequent kindergarten waves than on sample loss for the pre-kindergarten wave. In fact, while almost all (96 percent) of the cases affected by data breaches that agreed to participate in the kindergarten wave did so, 60 percent of the 181 ECLS-B cases that were affected by data breaches in the pre-kindergarten wave were lost to the kindergarten waves (109 cases).

2.6 Nonresponse Bias Analysis

The pre-kindergarten wave lost cases due both to nonresponse among the survey participants and to cases that were removed from the sample because of the data breaches. To better understand the effect these losses might have on estimates from the ECLS-B pre-kindergarten data, the population distributions of the nonresponding cases were computed across mother's age at child's birth, family structure, language of parent interview, mother's education, household poverty status, and urbanicity, and compared to the distributions for the full sample of cases eligible for the pre-kindergarten collection. Differences were confirmed using Chi-square tests of the distributions for the nonresponding cases relative to those in the full ECLS-B pre-kindergarten sample. Where differences occurred, the relative bias of each of the differences relative to the full sample was calculated and instances with relative bias over 10 percent were noted.

The only differences observed were by mother's education. There were relatively more mothers with a high school education or less among the nonresponders than there were in the full sample (26 vs. 19 percent for those with less than a high school education and 33 vs. 28 percent for those with a high school education). Conversely, there were relatively fewer mothers with more than a high school education in the population of nonresponding cases (40 vs. 52 percent).

The measurable effect of the data breaches on the pre-kindergarten data is relatively small. Overall only 1 percent of the pre-kindergarten sample was removed from the database because of the data breaches, and most of the key social and demographic variables had distributions that were not measurably different from those of the full sample.

However, the fact remains that approximately one-half of the cases affected by the data breaches were lost to the study, reinforcing the importance of avoiding data losses to data quality. A more detailed examination of the characteristics of the cases that were affected by the data breaches is a first step in learning about the decisions respondents made to stay or leave the study after being informed of the data breaches.

Looking first at the affected cases that chose to remain in the study, or at least to leave their data for the pre-kindergarten wave in the study, as was the case for survey nonrespondents, the one dimension where those who reconsented differ from the full sample is mother's education. However, among these cases the pattern is reversed from that observed among the nonrespondents. That is, a higher percentage of the mothers who agreed to stay in the study after being informed of the data breaches had more than a high school education (67 percent versus 52 percent in the full sample); conversely, the percentage of mothers who stayed in the study after being notified of the data breach was lower than the percent in the full sample for mothers with a high school education or less (20 versus 28 percent for those with less than a high school education and 12 versus 19 percent for those with a high school education).

Turning next to the cases that were removed from the database as a result of the data breaches, a different picture emerges. When these cases are considered as one group, the distributions differ from those of the full sample by mother's age at child's birth, family structure, language of parent interview, mother's education, and household poverty status. However, when these data are disaggregated further by those who actively refused to reconsent and those who were either not locatable or did not respond to the request for reconsent, the only comparison with a consistent pattern across the two groups is for mother's education, where there were relatively more mothers with less than a high school education and relatively fewer with more than a high school education in each category of the removed cases compared to the full sample. Those who were not located or did not respond also showed differences by family structure and

household poverty status compared to the full sample, with relatively more households without two parents and relatively more households below the poverty level. Those who actively refused the reconsent showed differences in language of the parent interview and urbanicity, with relatively more in the non-English group and in urban areas than there were in the full sample. Finally, while both groups whose cases were removed from the study showed differences relative to the full sample in the distribution by mother's age at child's birth, the differences were in opposite directions. Among those that actively refused to reconsent, there were relatively fewer cases in the 15 to 20 and 21 to 25 age groups and relatively more in the 26 to 30 and 31 and above age groups. In contrast, among those that were either not located or failed to respond to the reconsent request there were relatively more cases in the younger age groups (ages 15 to 20 and 21 to 25) and relatively fewer in age 31 and above age group.

3. Kindergarten Data Collection

3.1 ECLS-B Kindergarten 2006 Field Operations and Data Breaches[3]

A number of lessons learned from data breaches that occurred during the pre-kindergarten wave of data collection informed decisions made in the field procedures for the kindergarten 2006 data collection. In this wave, laptops were encrypted and strong passwords were required. Field interviewers continued using locked cases for both the laptop and other confidential information, and signature required FedEx priority overnight service was used for transmission of hardcopy materials. Field interviewers were required to transmit data for completed cases each day they worked, and an automatic logging function was used in the ECLS-B control system. The amount of confidential information included in the case folder was substantially reduced; the Address History Sheet continued to be used, but instead of the detailed Case Information Sheet, a Record of Action Form that only included household members' names was used to track contact attempts. Also, field interviewers used Daily Case Item Logs to track the location of all hard copy materials. As a result of these actions, there were fewer data breaches in the ECLS-B kindergarten data collection.

In addition, any actual or suspected loss of personally identifiable information was reported to the US CERT within one hour of NCES receiving notification. When loss was confirmed or attempts to locate missing items were unsuccessful, an Unanticipated Problem Report was prepared by the data collection contractor and sent to the NCES Project Officer, who in turn filed it with the NCES Chief Statistician, the ED CERT office, ED contracts office, and the NCHS ERB. The report described what was lost and proposed a risk-level assignment based on what was lost.

3.2 Data Breaches and the Assignment of Cases to Risk Categories

The losses that did occur were similar in nature to those described above for the pre-kindergarten wave of data collection. In total, there were 30 cases affected by data breaches that occurred as part of data collection during both rounds of the kindergarten wave of data collection (compared to 181 in the pre-kindergarten wave). Seventy percent of the data breaches during the kindergarten data collection involved the loss of an encrypted laptop but nothing else, and just under one-quarter (23 percent) involved the loss of personally identifiable information on paper forms in the field. Unrelated to the kindergarten data collection, an attempt to transfer videotapes from one contractor to another identified the loss of 19 videotapes from the data collections that were conducted in waves 1 and 2 of the study (9 months and 2 years). Since those losses were identified during the kindergarten data collection, they were included in the notification and reconsent process for the kindergarten wave and thus are included here in the analysis and description of the reconsent process and the effects on the resulting data.

Overall, 49 of the respondents to the kindergarten 2006 wave of ECLS-B (0.7 percent) experienced a data breach. Some 14 percent of these respondents (7 cases) experienced the loss of personally identifiable information and were offered credit monitoring. Because the amount of hard copy data used in the field was substantially reduced compared to the pre-kindergarten round of data collection, none of the respondents were placed at risk of having their survey data identified through a data breach. The remaining 42 affected respondents (86 percent) were considered to be at a minimal level of risk.

[3] At the time this paper was written, the kindergarten collection in 2007 had not been completed. Therefore, the experiences from the last round of data collection are not discussed.

When these 49 cases are considered as a group and compared to the full sample on a set of key social and demographic variables, the data show that relatively fewer children with mothers who were 21 to 25 years old at the time they were born and relatively more children with mothers who were 26 to 30 years-old at the time they were born experienced data breaches than expected based on the distribution by mother's age at child's birth in the full sample for the kindergarten 2006 wave.

Similarly, there were relatively more data breaches in urban areas and relatively fewer in rural areas than expected. In contrast, there are no significant differences in the distributions across the same set of variables between the cases without data breaches and the full sample.

3.3 Reconsent Results

One-half (51 percent) of the 49 cases affected by data breaches chose to stay in the study. Data for the remaining half of the affected cases were removed from the database. Thirty percent of the removed cases refused to reconsent, and 70 percent were not located. There were not measurable differences in the distribution of affected cases that chose to stay or had their data removed across the risk categories.

3.4 Resulting Survey Response Rates

The response rate for the cases that chose to stay in the study after being notified of data breaches was 80 percent (i.e., 5 out of 25 did not respond). The comparable response rate for the cases that did not experience data breaches in the kindergarten 2006 wave of data collection was 91 percent. The response rate for the subset of cases that chose to continue in the study after being notified of a data breach in the pre-kindergarten wave of data was 96 percent. When the results for those with and without data breaches reported in the kindergarten wave of data collection were combined, the unweighted response rate was 91 percent.

3.5 Nonresponse Bias Analysis

The kindergarten wave lost cases to nonresponse among the survey participants and to cases that were removed from the study because of the data breaches in the kindergarten wave of data collection. This is in addition to the cases that were removed from the study because of data breaches in the pre-kindergarten wave of data collection.

As a first step in understanding the effect the kindergarten losses might have on estimates from the ECLS-B kindergarten data, the population distributions of the nonresponding cases in the kindergarten data collection were computed using the same set of social and demographic variables that were used in the pre-kindergarten analysis. These data were then compared to the distributions for all cases that were eligible for the kindergarten data collection.

As was the case in the pre-kindergarten wave of data collection, the only differences in the distributions occurred by mother's education. There were relatively more mothers with a high school education or less among the nonresponders than there were in the full sample (26 vs. 19 percent for those with less than a high school education and 31 vs. 28 percent for those with a high school education). Conversely, there were relatively fewer mothers with more than a high school education in the population of nonresponding cases (42 vs. 53 percent).

The measurable effect of the data breaches on the kindergarten data is relatively small. One percent of the pre-kindergarten sample was removed from the database because of the data breaches, and another 0.3 of the pre-kindergarten sample was lost to the kindergarten wave as a result of affected cases leaving their pre-kindergarten data in the database but declining any further participation in the study. In addition, another 0.3 percent of the kindergarten sample was lost due to data breaches during the kindergarten 2006 collection. Combined, this amounts to less than 2 percent of the sample that started the pre-kindergarten wave of data collection.

However, as was the case in the pre-kindergarten wave of data collection, approximately one-half of the cases affected by the data breaches were lost to the study. Thus, to better understand who decides to stay in the study versus leave the study, and the effect the lost cases might have on the resulting data, the population distributions of these two groups of affected kindergarten cases were computed across a set of social and demographic variables and compared to the distributions for the full sample in the kindergarten data collection. In addition, having information on cases lost due to data breaches in two separate rounds of data collection provides the basis for a comparison of the distributions of the data breach cases across the two rounds, looking for similarities and differences in the characteristics of the cases lost to data breaches.

Looking first at the affected cases that chose to remain in the study, there are measurable differences between the full sample and the affected cases who remained in the study in the distributions by mother's age at the child's birth, household poverty status, and urbanicity. The affected cases that chose to remain in the kindergarten data collection after being notified of a data breach included relatively fewer children whose mothers were age 25 or younger at the time of their birth (12 versus 17 percent for the 15- to 20-year-old group and 12 versus 24 percent for the 21- to 25-year-old group) and relatively more children whose mothers were ages 26 to 30 when the ECLS-B child was born (44 versus 26 percent). There are relatively more affected cases who stayed in the study in households above the poverty level than in the full sample (84 versus 75 percent), and relatively more cases in urban areas (96 versus 85 percent).

The cases that were removed from the database as a result of the data breaches differ from the full sample for the kindergarten study only on mother's age when the child was born. In this case, the pattern is different from the pattern for the affected cases who remained in the study and somewhat different from the pattern observed among those who left the study in the pre-kindergarten wave. Compared to the kindergarten full sample, there were relatively more cases that left the study in the youngest age group (ages 15 to 20), relatively fewer in the 21 to 25 and 26 to 30 age groups, and then relatively more in the age 31 and above age group. Caution should be used in reading too much into the kindergarten results given the small numbers of cases involved in the data breaches.

4. Summary and Lessons Learned

Changes in field procedures between the pre-kindergarten and kindergarten data collections seemed to result in fewer instances of data breaches. All data breaches are regrettable and all measures possible should be taken to avoid them. Despite the emphasis in this report on the data breaches and the cases removed from the database due to data breaches, it is important to note that the across the two waves of data collection considered in this report a total of 230 cases experienced any data breaches, and these cases account for 2.5 percent of the eligible cases at the start of the pre-kindergarten data collection. In terms of the effect on the resulting database, the 150 cases that were lost as a result of the data breaches comprised 1.6 percent of the eligible cases at the start of the pre-kindergarten data collection.

Despite these small numbers, the analysis of the social and demographic characteristics of the cases removed from the ECLS-B database due to data breaches showed that in the pre-kindergarten wave there was differential loss from data breach cases removed from the database that were associated with mother's age at child's birth, type of family, language of interview, level of mother's education, household poverty status, and urbanicity. There was also differential loss in the kindergarten wave, with differences by mother's age at the child's birth, household poverty status, and urbanicity.

In both of these waves, the majority of cases were removed from the data collection for data breaches that were classified as negligible in the pre-kindergarten wave and lowest or minimal in the kindergarten wave, that is those cases where credit monitoring was not warranted. For example, 75 percent of the cases that were lost to the kindergarten wave of data collection due to the pre-kindergarten data breaches were in the lowest risk category, where credit monitoring was not warranted. In contrast, 25 percent of the lost cases were in the middle and highest risk categories and were offered credit monitoring. Similarly, in the kindergarten data collection, 82 percent of the cases that were removed from the study because of data breaches were due to either the loss of password protected encrypted laptops or to the loss of videotapes that included the child's and parent's images, but no last names or other personally identifiable information (no credit monitoring). By comparison, 14 percent of the cases that were removed from the study because of data breaches were in the group that had a data breach sufficient to warrant an offer of credit monitoring.

In May of 2007, the OMB Memorandum M-07-16 on "Safeguarding Against and Responding to the Breach of Personally Identifiable Information" included additional clarifications and guidance that are of particular relevance to federal data collections in light of the findings from this analysis of the data breaches in ECLS-B. In particular, OMB stated that "The likely risk of harm and the level of impact will determine when, what, how and to whom notification should be given." Especially noted is the fact that "Notice may not be necessary if, for example, the information is properly encrypted because the information would be unusable." Furthermore, because of the potential for a chilling effect from unnecessary notification, OMB indicated that "Agencies should exercise care to evaluate the benefit of

notifying the public of low impact incidents." and that "Agencies should bear in mind that notification when there is little or no risk of harm might create unnecessary concern and confusion."

If this guidance had been in place at the time of the ECLS-B pre-kindergarten and kindergarten data collections, NCES would have notified fewer cases and avoided raising undue concerns with many families. The number of notifications during the pre-kindergarten data collection might have been reduced from 181 to 41, and the number of cases removed from the study due to data breaches in the pre-kindergarten data collection would then have decreased from 109 to 27. By the same logic, the number of notifications during the kindergarten data collection might have been reduced from 49 to 7, and the number of cases removed from the study because of data breaches reported during the kindergarten data collection would then have decreased from 24 to 4.

Acknowledgements

The authors would like to thank the data collection staff at Research Triangle Institute (RTI) who provided the underlying data that made this analysis possible.