Research Report

Disclosure Risk in Educational Surveys: An Application to the National Assessment of Educational Progress

Andreas Oranje

David Freund

Mei-jang Lin

Yuxin Tang

Research & Development

June 2007
RR-07-24

Disclosure Risk in Educational Surveys:
An Application to the National Assessment of Educational Progress

Andreas Oranje, David Freund, Mei-jang Lin, and Yuxin Tang
ETS, Princeton, NJ

June 2007

Abstract

In this paper, a data perturbation method for minimizing the possibility of disclosure of participants' identities on a survey is described in the context of the National Assessment of Educational Progress (NAEP). The method distinguishes itself from most approaches because of the presence of cognitive tasks. Hence, a data edit should have minimal impact on both relations among demographic variables and relations between demographic and proficiency variables. Furthermore, since only a few students are at risk to be disclosed in a typical sampling setting common to educational surveys, the proposed data perturbation is governed by a nonuniform probabilistic process. The method is applied to data from NAEP and impact is computed using proficiency averages, demographic proportions, statistical inference results, and loglinear models. Results show that the proposed perturbation method has very little impact on NAEP results, even at relatively large editing rates. Some data coarsening results are reported as well. While the univariate results are relatively unaffected from the coarsening, loglinear models from higher order contingency tables are affected. It is recommended to restrict disclosure limitation techniques to perturbation methods in the case of NAEP.

Key words: Perturbation, NAEP, confidentiality edits, data coarsening, swapping

Acknowledgments

The work reported herein was in part supported under the National Assessment of Educational Progress (ED-02-CO-0023) as administered by the Institute of Education Sciences, U.S. Department of Education and by ETS allocation. The opinions expressed herein are solely those of the authors and do not necessarily represent those of Educational Testing Service. The authors are grateful to Dorota Staniewska for her help with the literature review and to Dan Eignor, Shelby Haberman, and Jiahe Qian for their reviews of earlier versions of this manuscript.

Table of Contents

Page

List of Tables

Page

List of Figures

Page

Introduction

Disclosure risk is the risk that someone from the general public can identify a participant of a survey by using online data analysis systems and linking individuals across data sources. In public surveys (e.g., the U.S. Census), the risk has to be limited to adhere to the law and to protect the privacy of all participants. Disclosure risk has been a relatively trivial problem in educational survey settings because publicly available data query tools and data analysis systems have been limited. The increasing focus on accountability over the past decade has changed this situation dramatically. Indicators of average proficiency in schools and school systems are prominently published in combination with detailed demographic distributions. Furthermore, an increasing level of interest is emerging in linking local information to state or national information, thereby creating a common ground for comparison. Hence, results of educational surveys are not only celebrating greater public attention, but also far more detailed scrutiny, motivating a rigorous evaluation of disclosure risks and possibly active reduction efforts. Prominent risks associated with disclosure are, besides legal provisions, discredited results in addition to declining participation (Fienberg & Willenborg, 1998), as privacy cannot be guaranteed. These considerations are obviously superseded by the notion that respondents' interests and right to privacy should be respected and protected.

An effective strategy that has been used to disclose respondents is to compare contingency tables of different dimensions (e.g., Boruch & Cecil, 1979, Chapter 7) and to attribute returning patterns across many combinations to individual test takers or otherwise sampled units. A subsequent step can then be to establish a link with other databases in an effort to learn about a host of attributes of the disclosed individual. Hence, limiting the disclosure risk has to be focused on reducing the usefulness of table comparisons with respect to individuals.

Many approaches to limiting disclosure risk have been developed (Dobra, Erosheva, & Fienberg, 2001; Fienberg, Makov, & Steele, 1998), including data perturbation, data coarsening, blanking or imputing, subsampling, adding random noise, combining extreme values of continuous variables into a single variable (top-bottom coding), and cell collapsing. Duncan and Pearson (1991) presented an elegant matrix representation of these methods as $\mathbf{Z} \rightarrow \mathbf{AZB} + \mathbf{C}$, where $\mathbf{Z}$ is the data matrix, $\mathbf{A}$ is the matrix that operates on cases, $\mathbf{B}$ is a matrix that operates on variables, and $\mathbf{C}$ is a matrix that adds noise. An alternative approach to disclosure risk limitation that has received much attention is the generation of "pseudo data" (Fienberg et al., 1998).

Fienberg and Makov (1998) and Dobra, Fienberg, and Trottini (2003) defined a Bayesian model that can be used to identify disclosure problems by generating populations through multiple imputation from a posterior distribution (following Rubin, 1987). Also, a clever algorithm has been proposed in which cumulative distribution functions are first smoothed and then used to sample pseudo data. A variant of this method is data shuffling (Muralidhar & Sarathy, 2003), which replaces sensitive data with simulated data that has similar distributional properties. The logic behind all of the above procedures is that even if an intruder believes he or she has identified a respondent, they cannot be sure that the information they obtained represents that respondent.

A large number of algorithms has been developed, often based on combinations of methods. Moore (1996) and Seastrom, Kaufman, Gonzales, and Roey (2004) provided detailed algorithms for data swapping using various distance measures to assess the impact of the swap. These algorithms are often based on a random selection of swap records and on swap partners obeying some kind of minimum distance measure and assuring that either marginal distributions (Dalenius & Reiss, 1982; Reiss, Post & Dalenius, 1982) or *t*-order ($t > 1$) marginal distributions (Reiss, 1984; Dobra, Karr, & Sanil, 2002; Fienberg & Slavkovic, n.d.) are preserved. Fienberg and McIntyre (2004) provided a post-randomization method to swap data in a range-restricted format. A crucial task is to appropriately define information content and information loss due to a decreased level of detail in categories or loss due to suppressed information caused by replaced values in the data. Gomatam and Karr (2003; see also Gomatam, Karr, & Sanil, 2004) compare several distance measures of the distortion in the joint distributions of categorical variables.

Swapping has certain advantages over the other disclosure risk reduction methods. It does not affect single variable, univariate analysis on the population and does not disturb the nonsensitive variables. However, there are also disadvantages. Specifically, data swapping might affect relationships between variables (e.g., correlations). Furthermore, since the swap rate is not disclosed so as to create uncertainty about whether a disclosure has nontrivial probability to be a true disclosure, users cannot incorporate the increased variability into their analyses. Perturbation methods can also cause inconsistencies in the data, which might be unattractive to statistical agencies and users (Fienberg & Willenborg, 1998). Despite these disadvantages, coarsening methods (e.g., suppression or cell collapsing) seem to carry many more problems than the methods just discussed. Suppression or collapsing is only effective if comparisons with higher or lower level tables are consistently suppressed or collapsed. Pannekoek and de Waal (1998) have developed an empirical Bayes method to determine whether a certain table or certain cells in the table, given the information that has already been disclosed in previous queries, pose a risk of disclosure. A substantial amount of bookkeeping is involved (Fienberg, 2001) and queries become dependent. In other words, the "first come, first serve" principle is adopted, which seems undesirable for results from a public educational survey. Nevertheless, there is a great deal of controversy regarding data distortion relative to suppression (e.g., Mackie & Bradburn, 2000, Chapter 4), which is mainly concentrated on the question of to what extent synthesized (e.g., swapping, imputation) data is useful to researchers and whether the synthesis procedures distort the kind of analyses that researchers perform and that are usually unknown to those who conduct the distortion.

An important consideration in choosing a method is the balance between risk and utility (Boruch & Cecil, 1979; Mackie & Bradburn, 2000; Trottini, 2003; Fienberg, 2001). The underlying notion is that disclosure risk can only be eliminated if the data are not released at all (Fienberg, 2001). The more data are published, the more the utility that is provided to the public and the higher the risk of disclosure is. The challenge is to find the point where the risk is acceptable without marginalizing utility, which varies from survey to survey. In this paper there will be substantial focus on higher order comparisons since univariate and bivariate results seem relatively straightforward to control under most swapping methods. Yet, it is not disclosure risk but rather the sample that may drive concerns about the disclosure of small cells. Specifically, highly detailed analyses may be questionable for a given sample. In addition, statistical considerations, such as the inability to establish a reliable mean or variance estimate, may drive the decision not to publish specific cells.

One of the most prominent examples of the application of disclosure limitation techniques involves the U.S. Census, using a combination of data swapping and coarsening. Depending on the type of data release, different approaches are employed (Zayatz, 2005). For the public use microdata samples, geographic and categorical thresholds (i.e., any identifiable area should at least have a population of 100,000 and any category at least 10,000), rounding prior to aggregation, noise addition, topcoding (categorization of continuous variables), and data swapping have been employed. For frequency count data, predominantly data swapping is applied in addition to thresholds. For magnitude data, cell suppression is used based on a percentage rule, which also includes complementary suppression as margins of the table are often available. There are exceptions for selected surveys where different or additional limitations are imposed.

In this paper, a data swapping algorithm for educational surveys is proposed. The educational survey situation is described first, followed by the algorithm. Additionally, a relative straightforward coarsening procedure is discussed. These methods are evaluated using real data from an educational survey and results are presented relative to unedited datasets. Finally, the results are discussed from practical, statistical, and program policy points of view.

Educational Surveys

A difference between most of the methods described in the aforementioned literature and educational surveys is the presence of cognitive data. In most applications such as census, health surveys, and genomic research (Lin, Owen, & Altman, 2004), uncertainty about summary statistics is predominantly attributed to sampling. For cognitive measures, a substantial part of uncertainty is driven by measurement imprecision, often expressed based on a measurement model (e.g., true score, item response theory). Because this imprecision appears at the individual level, disclosure risk of proficiency is generally not considered an issue. However, relations between variables that are considered to pose a risk (e.g., demographics) and proficiency should not be altered appreciably by data perturbation. In the case of assessments such as the National Assessment of Educational Progress (NAEP) demographic variables play an important role in the computation of proficiency and therefore proficiency becomes a concern for disclosure risk. This will be further discussed below.

NAEP is the only nationally representative and continuing assessment of what America's students know and can do in various subject areas such as reading, mathematics, and science. The assessment takes place in grades 4, 8, and 12 and for most academic subjects every two or four years. For this study, data of fourth- and eighth-graders on mathematics and reading have been used, containing records from all 50 states and the District of Columbia. According to Statistical Standard 4-2 of the National Center of Education Statistics (2002) files with personally identifiable information have to be edited through either coarsening, perturbation (e.g., swapping), or a combination of both.

Educational surveys have several characteristics in common that may alleviate some of the disclosure risk concerns. Foremost, these surveys are assessed in samples of students and therefore disclosure is only contingent on fairly detailed knowledge on the composition of the sample. Secondly, most background variables are based on self-reports and, therefore, are likely to contain a substantial number of errors compared to secondary records that are derived from parental reports, for example. Thirdly, and as mentioned before, there are several statistical considerations that call for caution in representing small cells. In NAEP, often statistics based on very few students are considered unreliable and are not reported. This is in addition to the fact that sample sizes (counts) are not reported at all. Finally, most educational survey programs draw stratified probability samples and apply several nonresponse weight class adjustments, leading to sampling weights. Identifying an individual based on weights poses a substantial difficulty as the relation between counts and weights is complex. Hence, the individual disclosure risk is relatively small to begin with and, therefore, relatively low impacting methods can be applied.

The NAEP model is based on the situation where students do not provide answers to all questions, but instead to a systematic subset. By pseudorandom assignment of subsets to students and by providing substantial overlap between items in subsets, a broad framework can be assessed in relatively short time. Consequently, there is little cognitive information gathered for individuals, and proficiency estimates based on individuals are inexact. The cognitive information can be aggregated across groups to obtain fairly precise estimates of proficiency distributions for groups, where groups are defined by any of the background variables. By writing this process in a regression form, proficiency is related to student groups as:

$$\theta = \boldsymbol{\gamma}\mathbf{x} + \varepsilon \tag{1}$$

where θ is the proficiency, γ is a vector of regression effects, *x* a vector of student group indicators represented as dummy-codes, and ε a residual term. Because proficiency is unobserved, a measurement model is assumed that relates observed responses *y* on cognitive items to proficiency following a specific function that is governed by parameters β and is expressed as the probability *P* of having a correct response. Furthermore, conditional on $\mu = \gamma x$, θ is assumed to be normally distributed and hence the marginal likelihood equation of the model is specified as:

$$L = \int_{\theta} P\left(\mathbf{y} \middle| \theta, \boldsymbol{\beta}\right) \varphi\left(\theta \middle| \mu, \sigma^2\right) d\theta \tag{2}$$

where φ is a normal distribution with mean μ and variance σ^2. Parameters of this model can be estimated and inferences about proficiency can be made.

The point of presenting Equations 1 and 2 here is to demonstrate how proficiency depends on the group definitions. A direct implication is that data perturbation should occur before scores are computed to avoid inconsistencies between the distributions of student groups and their proficiency distributions. In fact, it is conceivable that such an inconsistency can be used to reverse compute the swapping rate and the swapping variables and, subsequently, increase the level of disclosure risk. Nevertheless, proficiency will have to be an integral part of data perturbation to conserve relationships in the data. The option advocated in this paper is to use a proxy for proficiency, which is independent of student group distributions. More specifically, *normit* scores are used:

$$\tilde{\theta} = \Phi^{-1}\left(\tfrac{I}{I+J}\right) \tag{3}$$

where *I* is the number of correct responses, *J* the number of incorrect responses, and Φ^{-1} the inverse cumulative standard normal distribution. While this proxy can be considered quite coarse, the expectation is that it has sufficient detail for the current purpose.

Related to the notion that educational surveys are administered in samples of students is that risk of disclosure is higher for certain students. Specifically, if a certain student group is small (e.g., American Indians), then only few additional variables may be sufficient to disclose single cases. Hence, it seems prudent to select a swap sample using a nonuniform probability distribution to target specific at-risk groups. Obviously, the validity of this approach needs to be established.

Swapping Algorithm

What follows is a step-by-step description of the proposed swap algorithm. This procedure is easily adaptable to any educational survey.

1. Determine the swapping variables and swapping rate.
2. Form cells by concatenating the swapping variables. If, for example, 3 variables are used with 3, 4, and 2 levels, respectively, then 3 * 4 * 2 = 24 cells are formed.
3. Order the cells based on average normit score in each cell. This is the swap table.

4. Sample, using the a priori determined swapping rate, a subset that will be swapped. Students in typical small cells that are most at risk for disclosure will be assigned a higher probability of selection into the swap sample. A weight *u* will be assigned to each candidate with unit average that is based on the inverse probability of a student being in a cell of a k-way risk table composed of typical small group defining variables. Suppose that *u* is the swap probability weight and *R* is the swapping rate, then the probability of student *k* who is located in cell *g* of the risk table being selected in swap sample *S* is: $P\left(k \in S\right) = R \cdot u_k = \frac{n}{G \cdot n_g}$, where *n* is the total sample size (count), n_g the sample size (count) for cell *g*, and *G* the total number of cells in the risk table containing at least one student. These probabilities are computed without replacement and, hence, updated after every draw. No student sampling weights are used because the actual occurrence in a cell poses a risk, not the number of students in the population that are represented. The risk table is often composed of key reporting variables such as race/ethnicity, indicators of limited English proficiency, student disability status and school lunch eligibility.

5. Find the best possible swapping partner for each member of the swap sample from the two neighboring cells in the swap table. The best partner is the partner with normit closest to the normit of the member of the swapping subset.

   1. If the member of the swapping subset resides in the final or first cell of the ordering, then take the two adjacent cells before the end or after the beginning
   2. If one or both the cells are empty because either all members of the cell have already been used for swapping or are members of the swapping subset, then take the next cell in the ordering, unless it is the last or first cell, then take the next closest cell

6. From the two partners choose the partner with the least amount of swapping bias in terms of student sampling weights w (i.e., student group distributions). The swapping bias is represented as a distance measure and computed as:

$$d = \left(w_m \cdot \left(1 + \left|c_m - c_p\right|\right) + w_p\right) - \left(w_p \cdot \left(1 + \left|c_m - c_p\right|\right) + w_m\right)$$

where m is the member of the swapping subset, p a potential swapping partner, and c the number of the cell in the ordering. The absolute difference in c plus 1 is chosen to give the closest cell the advantage in case members are selected in cells at the beginning or end of the cell ordering.

7. Swap records.

Specifically for NAEP, because it is both a national and a state assessment, swapping occurs separately in strata (e.g., states) to assure that schools are not swapped across states and that marginal distributions of state-specific results remain unaltered.

Coarsening Procedures

The data perturbation method described here also has a coarsening component. As mentioned before, coarsening may require complex bookkeeping and is therefore undesirable. However, there may be practical situations that motivate the use of coarsening. For example, NAEP unperturbed results have been published for several decades and during that time access to the data was quite limited for the public. However, evolutions in online analysis tools allow anyone to query a database with individual records (see: www.nces.ed.gov/nationsreportcard). Although results are summarized before being displayed, this increased capability not only increases the disclosure risk for current assessments but also for all assessments that took place in the past. Yet, retrospective perturbation would invoke discrepancies between results published in the past and published under these new tools. Hence, coarsening might be a more desirable alternative.

The coarsening procedure applied in this paper is relatively straightforward. If a cell of a table is based on very few records, then the proportions of all the cells in the table are computed without those few records. Proportions in all other cells will be altered as well because the denominator of all proportions in the table has changed. While this method is relatively simplistic, the characteristics of educational surveys mentioned above (e.g., sample, weighting), in combination with the fact that most of the participants in past assessments are no longer in school, are deemed to provide sufficient disclosure risk reduction. The coarsening is set up to not alter the proficiency results. In other words, the proportion and the mean in a specific table may not be based on the exact same set of students, as the mean may include all students and the proportions may be adjusted for eliminated small cells.

In the following section the evaluation study will be described, followed by some results and a discussion.

Method

Both the coarsening and data swapping approaches described in the introduction are evaluated using data from the 2005 NAEP assessments. Data swapping is evaluated by swapping data from three states (California, Nevada, and New Jersey) at three different rates (0.5%, 1%, and 2.5% for effective swap rates of 1%, 2%, and 5%) in grade 8. Because evaluation of the swapping procedure is extensive, only one grade and a limited number of states were used with these states representing both homogeneous and heterogeneous populations with respect to the variables of interest. The swapping rates have been chosen for this study only and do not reflect actual swap rates that are used in operational NAEP. Results are compared to unaltered data. Specifically, the extent to which mean and proportion differences between student groups are significant or not across different swapping rates is evaluated. Furthermore, changes in statistical inference between 2003 (nonperturbed) and 2005 (perturbed at the 3 chosen swapping rates), compared to 2003 (nonperturbed) and 2005 (nonperturbed), are monitored as well. For this study, 4 out of 11 major reporting variables (e.g., student reported race/ethnicity, parental education) were swapped. These variables were chosen for this particular study and do not represent variables that were used in swapping for operational NAEP. The variables and swap rates used in operational NAEP are not publicly available in an effort to further reduce the disclosure risk.

Data coarsening is evaluated across all jurisdictions (all states and Washington, D.C.) for grades 4 and 8 in 2005 Reading and Mathematics. Besides monitoring the number of times cells in one-, two-, and three-way tables are coarsened following the coarsening procedure described above, changes to subgroup proportions also are computed. Furthermore, using some primary reporting variables (i.e., Race/Ethnicity, Individualized Education Plan Status, English Language Learner Status, Student Reported Parental Education), loglinear models are computed before and after coarsening for all jurisdictions in grade 8 Reading. Loglinear models were chosen as a typical analysis that a user may conduct on NAEP’s demographic variables with distributional results (specifically, weighted percentages) retrieved from the publicly accessible online data analysis system NAEP Data Explorer (NDE). Other types of analyses on weighted percentages are certainly conceivable. Incidental zeros are replaced with 0.001. The number of incidental zeros was quite large, indicating structural voids. However, since the interest here is not a specific model but the impact of coarsening, no jurisdiction-specific model adjustments were made. Ad hoc investigation of the likelihood ratio test (LRT) with and without zero adjustment revealed that the change in LRT is trivial (less than 0.5% relative to the degrees of freedom), especially compared to the effect of coarsening (up to 15% relative to the degrees of freedom). Four different models are computed: full factorial, up to all three-ways, up to all two-ways, and a main effects only model. Before- and after-coarsening results are compared with respect to parameter estimates, significance of test statistics, and fit indices of the model.

Results

Swapping and coarsening are initially discussed separately. All comparisons are made in terms of the impact of these procedures on the results.

Swapping

Swapping is evaluated as the difference between the swapped and unswapped data in terms of current NAEP statistics of interest such as mean scale scores, achievement level percentages, and student group percentages. For each of the groups, these statistics are computed and, subsequently, the deviation of the unswapped results from the swapped results (unswapped minus swapped) divided by the standard error of the unswapped results is obtained. Figures 1 through 3 show the average difference in all three states, both across *all* variables and then only those variables that were swapped. Note that under the NAEP model proficiency results depend to some extent on the distributions of demographic variables and, therefore, proficiency estimates for *all* students may change due to swapping. Figure 1 addresses scale score means, Figure 2 the percentage at or above Proficient (from three levels: Basic, Proficient, and Advanced), and Figure 3 student group percentages.

From Figures 1 and 2 it becomes clear that the swapping procedure has little influence on the average scale scores or achievement level percentages. Also, there is no specific pattern discernible across the swapping rates. Furthermore, proficiency estimates for subgroups defined by swapped variables seem not to be more affected than estimates for subgroups based on nonswapped variables. With respect to the student group percentages, a fairly strong pattern of swapping rate and impact was found for the states New Jersey and Nevada. However, the size of

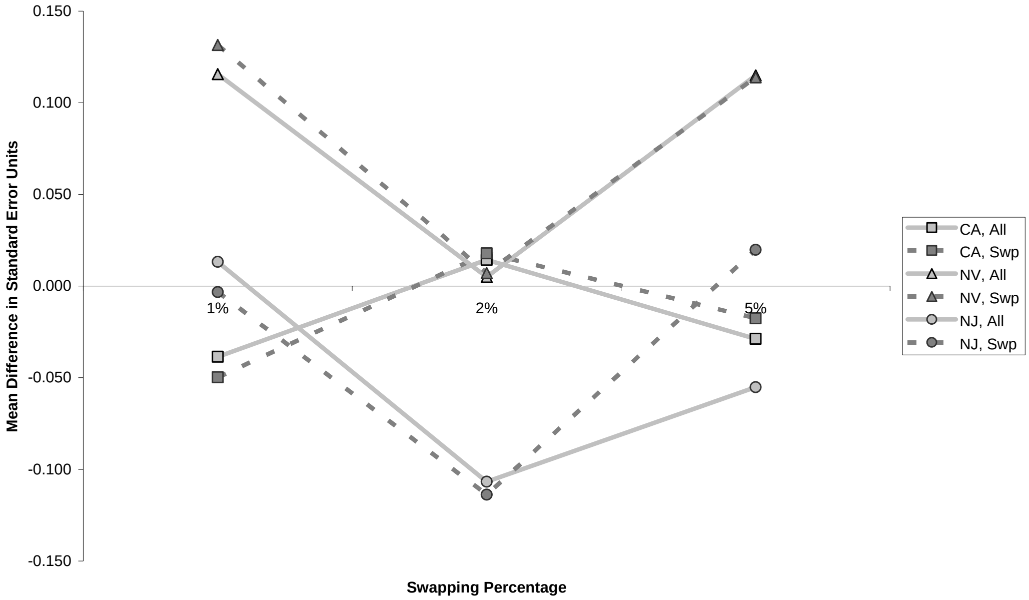

Figure 1. **Mean difference in standard error units of the mean score, all reporting variables (all) and swapped variables only (swp).**

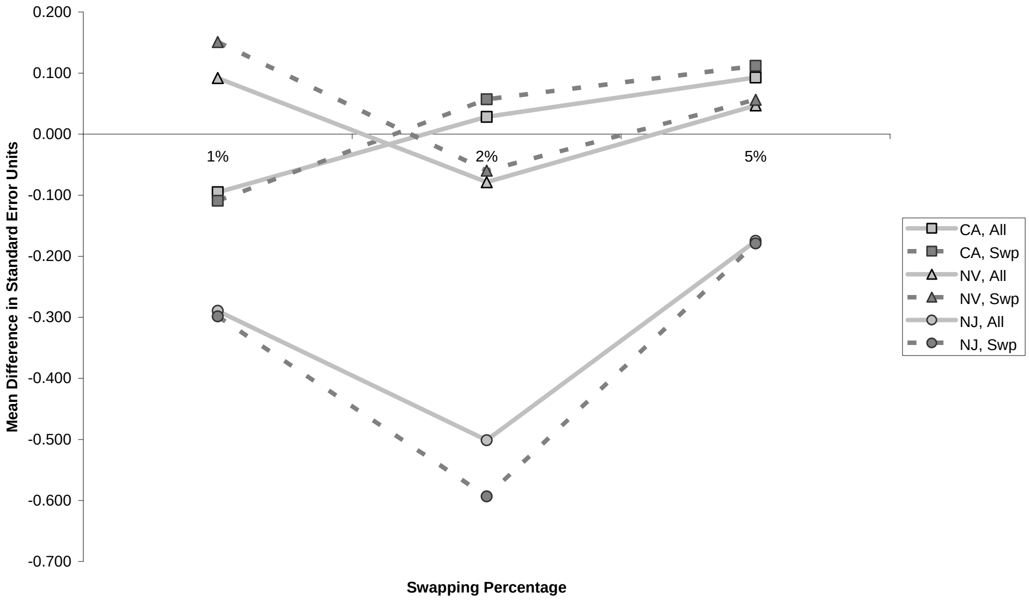

Figure 2. **Mean difference in standard error units of the percentage at or above proficient, all reporting variables (all) and swapped variables only (swp).**

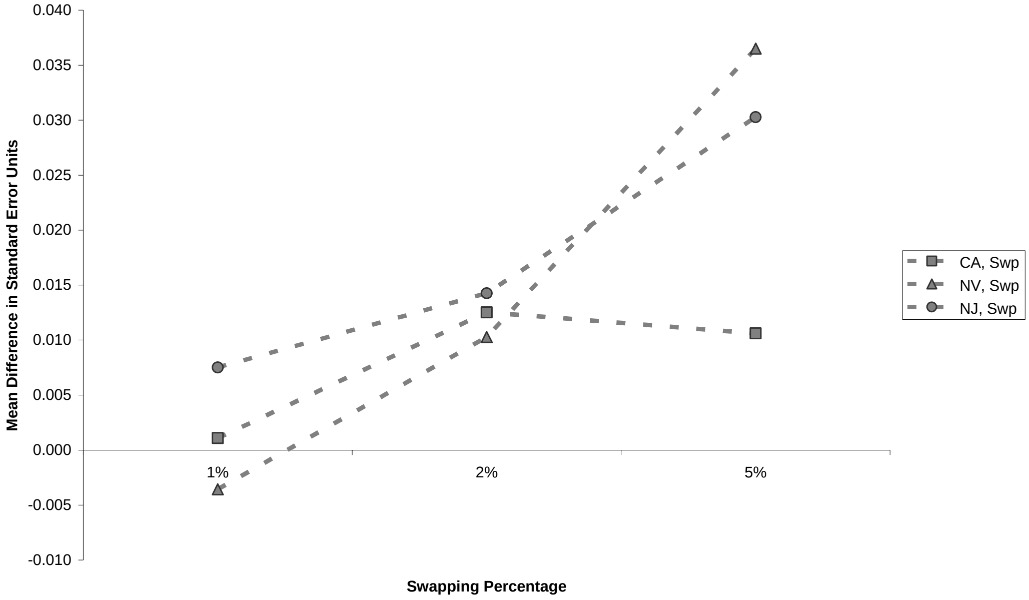

Figure 3. **Mean difference in standard error units of the student group percentage, swapped variables only (swp).**

the impact is trivial. Note that only the swapped variable averages are displayed since the unswapped variables are not affected by the definition of the procedure. Figures 4 through 6 address the same statistics except that the bias in the standard error (unswapped minus swapped) is displayed. Similar to the average statistic differences, no particular pattern is discernible for either the standard error of the mean scale score or the percentage at or above proficient. Likewise, a similar pattern is found for the student group percentages, with an increase in the variance as the swap rate goes up. Yet, the bias appears to be marginal for all three states.

From the results presented here it is concluded that the swapping procedure has a minimal effect on the univariate results. It should be noted that only one swapping was performed, which limits the generalizability of these results. Also, no multivariate results were inspected.

Coarsening

Coarsening was evaluated in two ways. First, the impact of coarsening on student group percentages was assessed for univariate tables. Subsequently, loglinear models were estimated both with and without coarsening to assess the impact of coarsening.

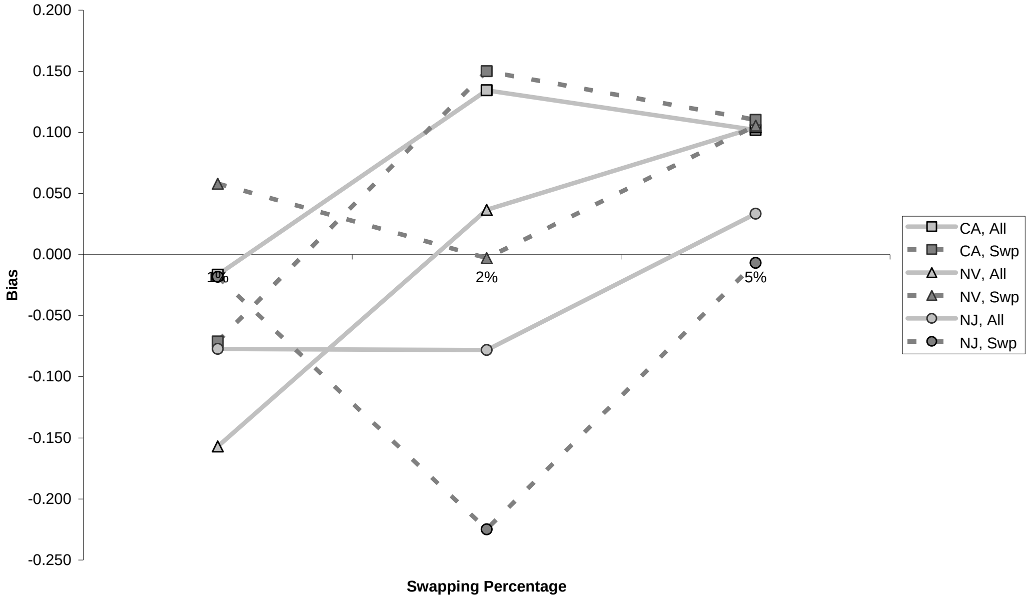

Figure 4. **Bias in standard error of the mean score, all reporting variables (all) and swapped variables only (swp).**

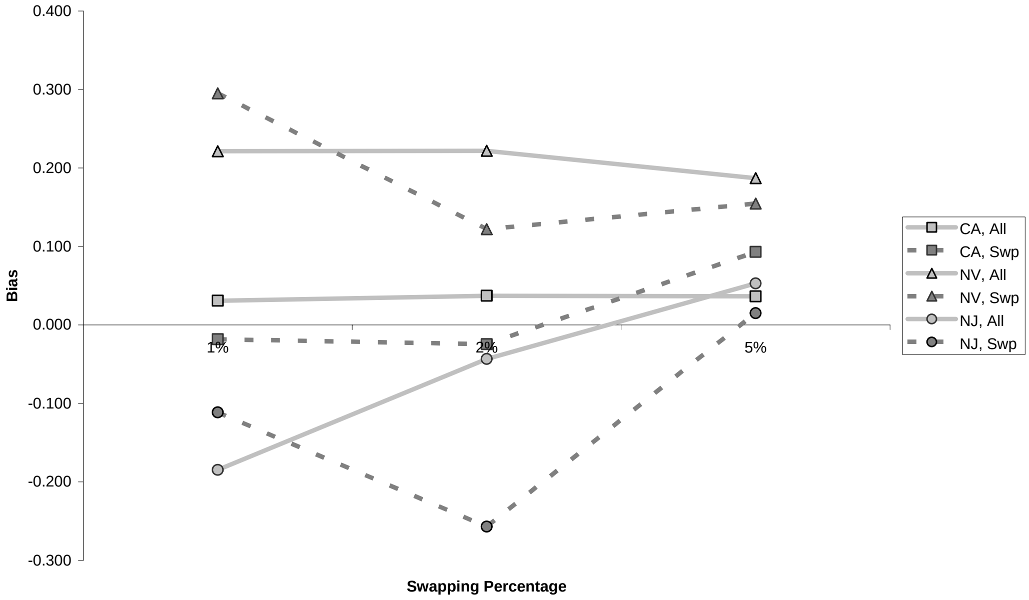

Figure 5. **Bias in standard error of the percentage at or above proficient, all reporting variables (all) and swapped variables only (swp).**

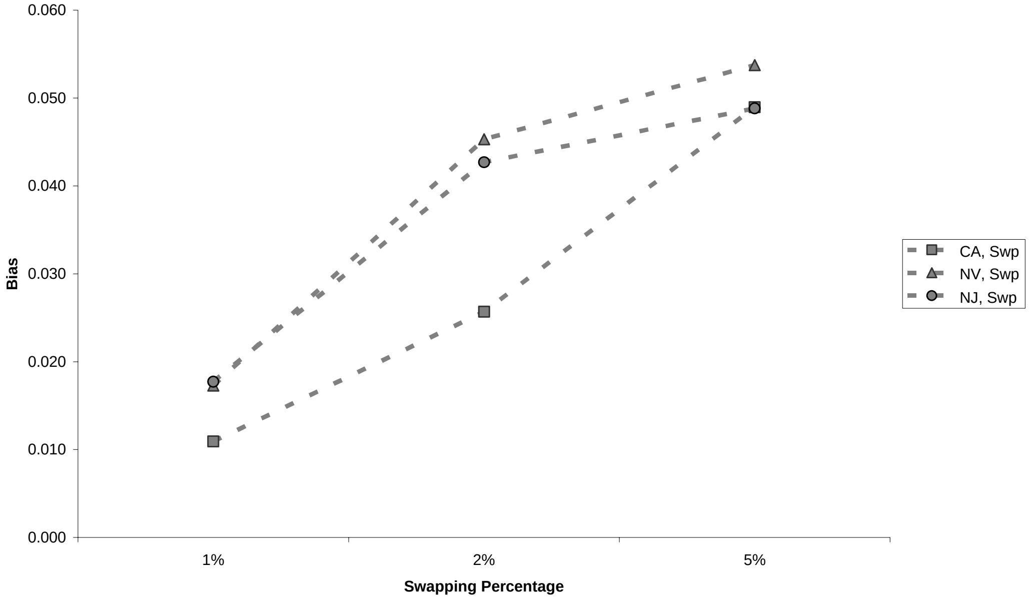

Figure 6. **Bias in standard error of the student group percentage, swapped variables only (swp).**

Across all 50 states and the District of Columbia, the 11 key reporting variable student group percentages were computed both with and without coarsening, for grades 4 and 8 Reading and Mathematics, and for 2003, 2005, and the difference between 2003 and 2005. Subsequently, the percentage was divided by the standard error of the without-coarsening result and a table was constructed to assess whether the results of both with and without coarsening were consistent in terms of significance (absolute value of a t-statistic greater or smaller than 1.96). Table 1 provides the inconsistency percentage. Also, the mean and the maximum positive and negative differences between coarsened and uncoarsened percentages for key student groups across all states were computed. To give an indication of the importance of these means, NAEP results of major reporting groups in New Jersey, Nevada, and California in grade 4, 2005 can be consulted. Small standard errors of 0.1 are associated with the percentage of American Indians in New Jersey and California. Relatively large standard errors for student group percentages are 2.5 (not eligible for free lunch in New Jersey) and 1.8 (central city in Nevada).

Table 1 shows that the inconsistency percentage is trivial. Table 2 shows that the average impact of the coarsening procedure is small, but that locally, relatively large differences can be found of up to 3 percent. However, the standard errors associated with these differences are probably large as well given the trivial inconsistency percentages in Table 1.

Figures 7 through 10 show the average difference in percentage between coarsened and uncoarsened data for both grades and subjects for specific variables. These variables are Race/Ethnicity (White, Black, Hispanic, Asian, American Indian, Other), English Language Learner (ELL) Status (Yes, No, Formerly), and school Lunch Eligibility (Eligible, Not Eligible, No Information Available), chosen because these are frequently subject to coarsening. The figures show that the larger percentages (e.g., Whites) are most affected by the coarsening, which can be easily shown mathematically. However, some of the smaller groups, such as American Indians, are subject to change as well because these categories are predominantly subject to coarsening themselves, changing their proportions to zero.

Table 1

Inconsistency Percentage Between Coarsened and Uncoarsened Student Group Percentage Significancies of NAEP Grades 4 and 8 Reading and Mathematics for 2003, 2005, and the Difference Between 2003 and 2005 Across All States for 11 Key Reporting Variables

Subject	Grade	2003	2005	2005-2003
Mathematics	4	0%	0%	0%
	8	0.08%	0.09%	0.26%
Reading	4	0.10%	0%	0.34%
	8	0%	0%	0.18%

Note. N of cells is between 893 and 1,278.

Lastly, loglinear models were estimated using four variables: Student Reported Parental Education, Race/Ethnicity, ELL status, and Individualized Education Plan (IEP) status. Three different models were estimated before and after coarsening: full factorial, all-three way effects and below, all two-way effects and below, and main effects only. The models were run in all 51 jurisdictions in grade 8 2005 Reading . On average, 44.2% of the cells were coarsened, where the total number of valid cells (at least one observation) was on average 66. As a result, on average, 187 cells are empty and, as mentioned before, the number of observations for purposes of loglinear modeling is set to 0.001 in these cells.

Table 2

Mean and Maximum Positive and Negative Differences Between Coarsened and Uncoarsened Student Group Percentages of NAEP Grades 4 and 8 Reading and Mathematics for 2003, 2005, and the Difference Between 2003 and 2005 Across All States for 11 Key Reporting Variables

Subject	Grade	Statistic	2003	2005	2005-2003
Mathematics	4	Mean	-0.007	-0.003	0.001
		Max. positive	0.186	0.150	0.799
		Max. negative	-1.776	-1.726	-0.275
	8	Mean	-0.012	-0.012	0.001
		Max. positive	0.455	0.217	1.323
		Max. negative	-2.667	-2.811	-0.946
Reading	4	Mean	-0.017	-0.010	-0.003
		Max. positive	0.338	0.141	0.785
		Max. negative	-3.108	-1.776	-1.275
	8	Mean	-0.008	-0.012	0.002
		Max. positive	0.382	0.177	0.951
		Max. negative	-1.872	-1.750	-0.799

Note. *N* of cells is between 893 and 1,278.

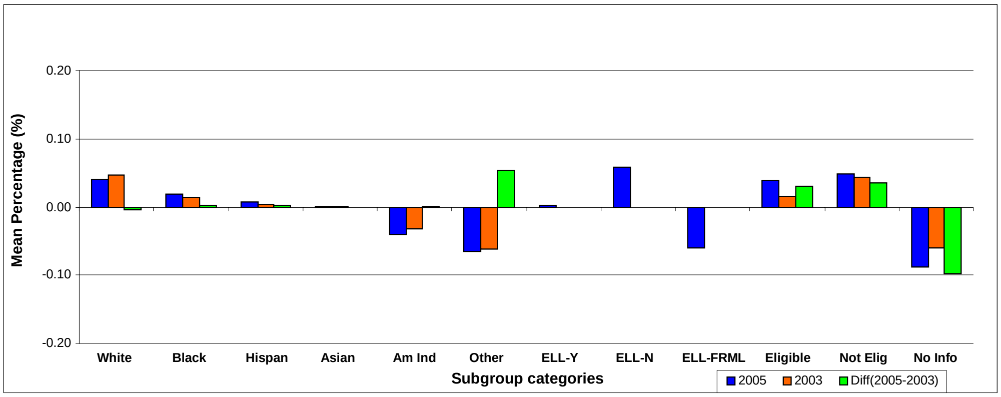

Figure 7. **NAEP Math 2005 Grade 4 mean percentage differences between coarsening and noncoarsening among states by subgroup.**

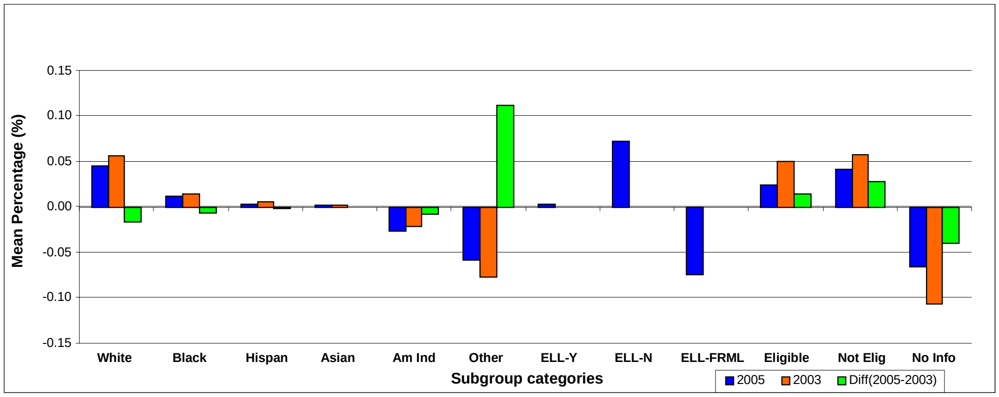

Figure 8. **NAEP Math 2005 Grade 8 mean percentage differences between coarsening and noncoarsening among states by subgroup.**

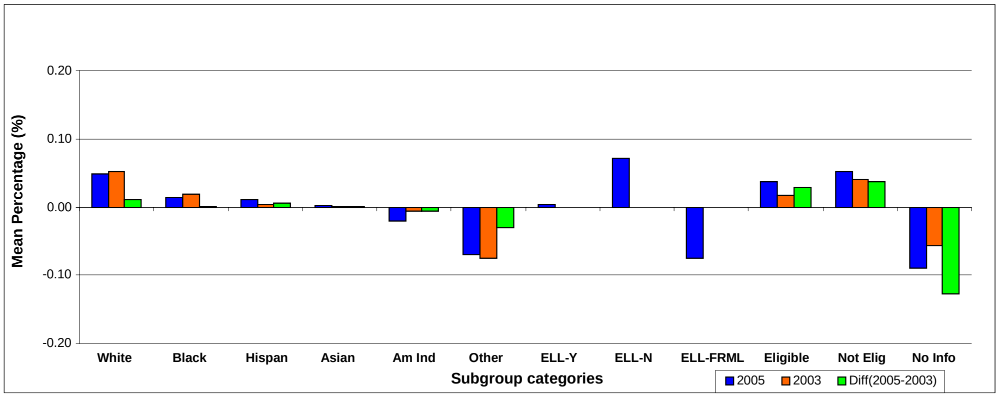

Figure 9. **NAEP Reading 2005 Grade 4 mean percentage differences between coarsening and noncoarsening among states by subgroup.**

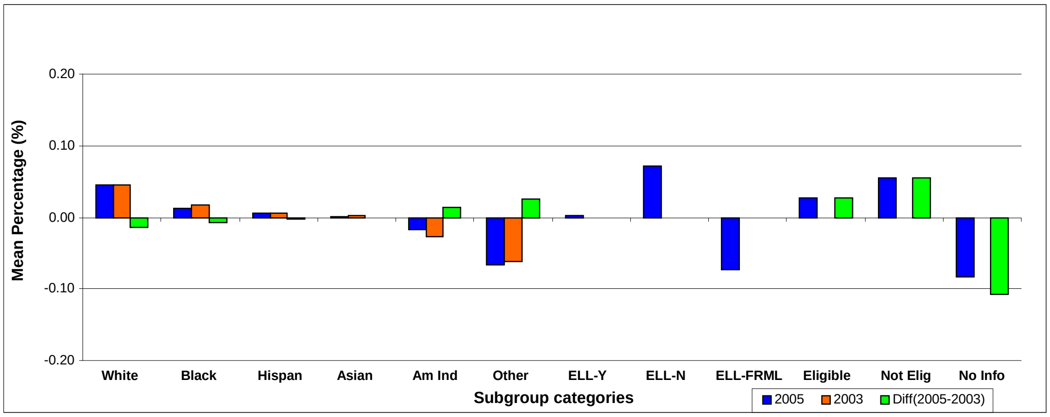

Figure 10. **NAEP Reading 2005 Grade 8 percentage differences between coarsening and noncoarsening among states by subgroup.**

The significance of the loglinear effects was determined through a chi-square test and compared between the coarsened and the noncoarsened data. Across the four models, 47 parameters are estimated in each of 51 jurisdictions. Of those 2,397 effects, 2,287 or 95.4% are consistent between the coarsened and noncoarsened data, 67 or 2.8% are significant in the noncoarsened data but not significant in the coarsened data, and 43 or 1.8% are significant in the coarsened data but not significant in the noncoarsened data. Table 3 shows the inconsistencies by model revealing that the two-way models are relatively inconsistent compared to the other models. No adjustment for multiple comparisons was conducted to reflect a conservative inquiry with respect to detecting consistencies.

Table 3

Inconsistency Rates by Model of the Parameter Estimate Significancies From Zero Between the Uncoarsened and Coarsened Data

	Noncoarsened significant, coarsened not significant	Coarsened significant, Noncoarsened not significant
Full factorial	1.6%	0.9%
Up to three way	2.5%	2.4%
Up to two way	6.1%	3.2%
Main effects only	0.4%	0%

Table 4 shows the differences between the coarsened and noncoarsened data-based parameter estimates for each of the models averaged over jurisdictions and expressed in standard errors units of the model parameters based on the noncoarsened data. The table shows that the impact of coarsening is substantial for the main effect of race/ethnicity in all four models except the full factorial model, being at least half a standard error unit. In the main effects only model the intercept and ELL status are also substantially impacted. In the two-way model, besides race/ethnicity, the interaction of race/ethnicity with parental education and IEP status also shows nontrivial change due to coarsening.

Table 4

Differences Between the Coarsened and Noncoarsened Data-Based Parameter Estimates for Each of the Four Models Averaged Over Jurisdictions and Expressed in Standard Errors Units of the Model Parameters Based on the Noncoarsened Data

	Full factorial	Up to three way	Up to two way	Main effects
Intercept	0.107	0.141	0.353	-1.139
A[a]	-0.425	-0.668	-2.414	-2.558
B[b]	-0.042	-0.058	-0.419	0.468
C[c]	0.129	0.204	0.416	1.417
D[d]	-0.008	-0.014	-0.028	0.360
A * B	0.213	0.341	1.821	
A * C	0.010	-0.014	0.248	
A * D	0.052	0.092	0.529	
B * C	-0.101	-0.170	-0.114	
B * D	-0.017	-0.029	-0.098	
C * D	0.005	-0.002	-0.063	
A * B * C	0.056	0.184		
A * B * D	-0.021	-0.071		
A * C * D	-0.002	-0.008		
B * C * D	-0.009	0.008		
A * B * C * D	0.016			

[a] Race/ethnicity as reported by the school. [b] IEP status. [c] ELL status. [d] Parental education as reported by the student.

In sum, while a substantial number of cells were coarsened in the examples presented here, the impact of this coarsening was relatively small except for the lower order models where parameter estimates differ substantially for certain variables.

Conclusion and Discussion

In this study two disclosure risk limitation techniques are described, developed and evaluated in the context of NAEP. Disclosure risk is the risk that someone from the general public can identify a participant of the NAEP survey by using online data analysis systems such as the NDE. The risk has to be limited to adhere to the law and to protect the privacy of all participants. Two approaches are applied to NAEP: data coarsening and data swapping. In short, the data is coarsened by removing the results from all cells that are based on either 1, 2, or 3 observations in a table. This also impacts the denominator of all other cells. Data swapping entails the random swapping of certain records with respect to a few to the public unknown variables. Under this scheme claims about the identification of a specific individual cannot be substantiated due to the probability that the record of interest might have been swapped. For NAEP, a specific swapping algorithm was developed that targets cells that are particularly at risk for disclosure, to be included with relatively high probability in the swap sample.

The results indicate that data swapping has a relatively small impact on both univariate and multivariate results (i.e., student group percentages and scale score means). The impact is somewhat larger in small student groups, yet well below any notion of substantial differences. Data coarsening seems to be relatively harmless for univariate trend results (i.e., student group percentages). However, data coarsening has some impact on parameter estimates of loglinear models from multivariate contingency tables. Hence, it is advised to refrain from coarsening in NAEP and to limit the disclosure limitation to data swapping type approaches.

The results presented in this paper should be placed in context. For example, only a few variables were evaluated and it is quite possible that other variables are more affected by data perturbation. Also, only a subset of jurisdictions was used in the swapping evaluation and for one year only. Results from national samples, as opposed to state samples, and from multiple years might be less or more vulnerable to data swapping. Also, the NAEP sample has several unique characteristics and the scale scores have a particular form that makes generalizations possibly prohibitive.

In closing, it should be reiterated that the only completely successful disclosure risk limitation technique, is to not release data at all. Hence, the legal impetus and requirement is in some sense unattainable in a practical setting where public funding suggests that public data utility is provided. NAEP is in the circumstance where disclosure risk is minimal to begin with due to the characteristics of the survey: a sample is drawn, not everyone receives the same academic subject, no individual answers enough cognitive items to warrant a reliable individual scale score, and sampling weights dilute a straightforward relationship between percentages and the number of students. Furthermore, neither student counts nor results for smaller student groups are reported, where the statistical reliability of the standard error of the estimate is known to be low or where the estimate cannot be ascertained. Hence, the approaches discussed in this paper have minimal impact on the accuracy of the results.

References

Boruch, R. F., & Cecil, J. S. (1979). *Assuring the confidentiality of social research data.* Philadelphia: University of Pennsylvania Press.

Dalenius, T., & Reiss, S. P. (1982). Data swapping—A technique for disclosure control. *Journal of Statistical Planning and Inference 6*, 73–85.

Dobra, A., Erosheva, E. A., & Fienberg, S. E. (2001). *Disclosure limitation methods based on bounds for large contingency tables with applications to disability.* Retrieved May 1, 2007, from http://www.stat.washington.edu/adobra/Files/Papers/TEN-DEF-11-09-02.pdf

Dobra, A., Fienberg, S. E., & Trottini, M. (2003). Assessing the risk of disclosure of confidential categorical data. In J. M. Bernardo, M. J. Bayarri, A. P. Dawid, J O. Berger, D. Heckerman, A. F. M. Smith, et al. (Eds), *Bayesian statistics 7: Proceedings of the seventh Valencia international meeting* (pp. 125–144). Oxford, UK: Oxford University Press.

Dobra, A., Karr, A. F., & Sanil, A. P. (2002). Preserving confidentiality of high-dimensional tabulated data: statistical and computational issues. *Statistics and Computing, 13*(4), 363-370.

Duncan, G., & Pearson, R. (1991). Enhancing access to data while protecting confidentiality: prospects for the future. *Statistical Science, 6*(3), 219–239.

Fienberg, S. E. (2001). Statistical perspectives on confidentiality and data access in public health. *Statistics in Medicine, 20*, 1347–1356.

Fienberg, S. E., & McIntyre, J. (2004). Data swapping: Variations on a theme by Dalenius and Reiss. In J. Domingo-Ferrer & Vi. Torra (Eds.), *Lecture notes in computer science*: *Vol. 3050. Privacy in statistical databases: PSD 2004 Proceedings* (pp. 14–29) New York: Springer-Verlag.

Fienberg, S. E., & Makov, U. E. (1998). Confidentiality, uniqueness and disclosure limitation for categorical data. *Journal of Official Statistics, 14,* 385–397.

Fienberg, S. E., Makov, U. E., & Steele, R. J. (1998). Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics, 14,* 485–502.

Fienberg, S. E., & Slavkovic, A. B. (n.d.). Making the release of confidential data from muti-way tables count. Retrieved September 1, 2004, from http://www.niss.org/dgii/techreports.html

Fienberg, S. E., & Willenborg, L. C. R. J. (1998). Introduction to the special issue: Disclosure limitation methods for protecting the confidentiality of statistical data. *Journal of Official Statistics, 14*, 337–345.

Gomatam, S., & Karr, A. (2003). *Distortion measures for categorical data swapping* (Technical Rep. No. 131). Research Triangle Park, NC: National Institute of Statistical Sciences.

Gomatam, S., Karr, A. F., & Sanil, A. P. (2004). *Data swapping as a decision problem.* Retrieved September 1, 2004, from http://www.niss.org/dgii/techreports.html

Lin, Z., Owen, A. B., & Altman, R. B. (2004). Genomic research and human Ssbject privacy. *Science, 305*, 183.

Mackie, C., & Bradburn, N. (Eds.). (2000). *Improving access to and confidentiality of research data: Report of a workshop.* Washington, DC: National Research Council, National Academy Press.

Moore, R. A. (1996). *Controlled data-swapping techniques for masking public use microdata sets* (Statistical Research Division Report Series, RR96-04*)*. Washington, DC: U.S. Bureau of the Census.

Muralidhar, K., & Sarathy, R. (2003, April). *Masking numerical data: Past, present, and future.* Presentation to the Confidentiality and DataAccess Committee of the Federal Committee on Statistical Methodology, Washington DC.

National Center for Education Statistics. (2002). *2002 NCES Statistical standards, Standard 4-2.* Retrieved on September 22, 2004, from: http://nces.ed.gov/statprog/2002/std4_2.asp

Pannekoek, J., & de Waal, A.G. (1998). Synthetic and combined estimators in statistical decision control. *Journal of Official Statistics, 14*, 399–410.

Reiss, S., Post, M., & Dalenius, T. (1982). Non-reversible Privacy Transformations. In *Proceedings of the first ACM SIGACT-SIGMOD symposium on principles of database systems*, 139–146. New York: ACM Press.

Rubin, D. B. (1987) *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Seastrom, M., Kaufman, S., Gonzales, P, & Roey, S. (2004). Do disclosure controls to protect confidentiality degrade the quality of the data? United States Data from the 2002 Trends in Mathematics and Science Study. Paper presented at the European Conference on Quality and Methodology in Official Statistics (Q2004). Mainz, Germany.

Trottini, M. (2003). *Assessing disclosure risk and data utility: A multiple objectives decision problem.* Retrieved September 22, 2004, from http://www.unece.org/stats/documents/2003/04/confidentiality/wp.19.e.pdf

Zayatz, L. (2005). *Disclosure risk avoidance practices and research at the U.S. Census Bureau: An update* (Research Rep. Series No. 2005-06). Washington, DC: U.S. Census Bureau.