Expert Task Force Consideration of Confidentiality, Effect Sizes, and Computerized Adaptive Testing at NCES

White Mitchel and Mariana Watson

National Center for Education Statistics
University of Madrid

Abstract

The National Center for Education Statistics (NCES) has established a relationship with the National Institute of Statistical Sciences (NISS) to help it review its data collection activities (mainly by forming and supporting task forces of recognized experts in methodology and data collection) and produce findings regarding best practices. This paper summarizes the findings of task forces on three topical areas and related NCES activities. The three topic areas discussed in this paper are:

- Making data accessible while protecting student and school confidentiality;
- The usefulness of reporting effect sizes in NCES publications; and
- The feasibility and potential value of using Computerized Adaptive Testing (CAT) for assessments in NCES longitudinal studies.

For a complete list of topics NISS has helped NCES address or is helping the agency address see Appendix 1 - Other NISS Work for NCES.

Background

The Organizations

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting statistics and information showing the condition and progress of education in the U.S. in order to promote and accelerate the improvement of American education. NCES is located within the U.S. Department of Education and the Institute of Education Sciences. As well as providing data to federal policy makers, part of the NCES mission is to serve the research, education and other interested communities. NCES strives to adopt and develop the best possible statistical and survey research practices in its work.

The National Institute of Statistical Sciences (NISS) was established in 1991 by the national statistics societies and the Research Triangle universities and organizations to identify and foster high-impact, cross-disciplinary research in the statistical sciences. NISS serves the statistical sciences communities by: performing research at the interface between statistics and disciplinary science, as well as between industry/government and academia; supporting career development at all levels, with special emphasis on postdoctoral fellows; and engaging the national statistical sciences community in a variety of activities.

 ISSN 2047-3745

The Mechanism
In the early 1990s NCES' increasing demand for contractor support of its mission resulted in a long-term contract with AIR, Inc. to establish the Education Statistics Services Institute (ESSI). First formed in 1995 (and re-awarded in 2005) ESSI is a consortium of twelve organizations that supports the programs of NCES. NISS operated at first under sub-contract to ESSI and later joined the consortium. http://www.air.org/essi/essi_main.aspx

This partnership allows NCES to meet its ongoing needs for expert consultation with academic and industry on cutting-edge statistical and survey research topics to improve and evaluate its data collection activities. The needs have been successfully and productively met by using NISS to help identify and secure the cooperation of relevant experts, bring them into topic-specific groups along with NCES experts, and produce informative topical reports grounded in the framework of best theory and practice.

Task Force on Confidentiality

Issue
Making data accessible while protecting student and school confidentiality.

Charge
The principal goals of the NISS task force on confidentiality were to review the NCES current (2002) and planned data dissemination strategies for confidential data, assessing whether these strategies were appropriate in terms of both disclosure risk and data utility, and then to suggest potential changes that current methodology may allow. (Karr A., NISS/NCES Data Confidentiality Task Force: Final Report, 2008).

Membership
See Appendix 2 - Task Force Membership.

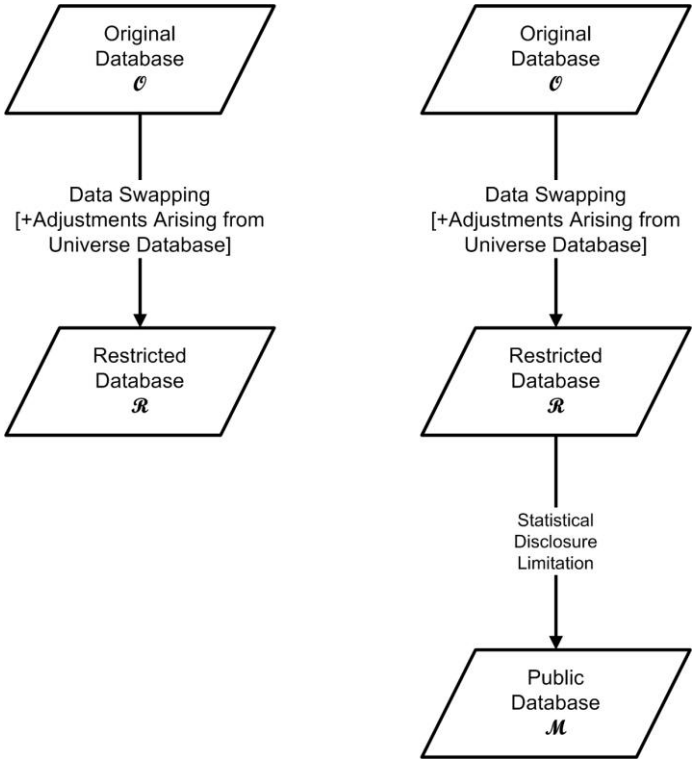
Problem Formulation
Common to all situations considered, and shown in the left-hand panel in Figure 1, are:

1. An *original database* O, as collected and edited (for instance, to adjust for nonresponse bias) by an NCES contractor.
2. A *restricted database* R, produced by the data collection contractor from the original database, using the NCES DataSwap software, designed to be accessible only by license through NCES. R may also be adjusted to maintain consistency with associated universe databases.
3. A public database, M (for "masked"), produced from R, by application of one or more methods for statistical disclosure limitation (SDL), which is available to the public without licensing or any other restriction.

Figure 1: Formulation of the dissemination problem

System without a public database System with a public database available without any licensing or other restrictions

 ISSN 2047-3745



Each of R and M (if the latter exists) could potentially be accessible in two conceptually distinct ways:

1. Directly, in the case of R by obtaining a copy under license from NCES, and in the case of M by downloading a copy from an NCES web site. Any statistical analysis may be performed on either R or M.
2. Electronically, by means of online data access systems (DASs), to which users submit queries specifying statistical analyses to be performed on R or M.

The task force understood that NCES is committed to access by license to R in all cases, and was eager to provide DAS access to R and/or M if confidentiality were not threatened. Consequently, for each NCES data collection, three decisions are necessary:

1. whether and under what circumstances to allow online DAS access to R, as well as the nature of such access;
2. whether to produce and make available a public database M; and,
3. if there is a public database M, whether and under what circumstances to allow online DAS access to it, as well as the nature of the access.

Task Force Findings and Suggestions

Overall. Continue to treat the restricted database R as "ground truth" in the sense that all NCES analyses and publications are based on it rather than on O. This ensures consistency between internal and external analyses.

Transformation of O to R. Employ the DataSwap software to create R from O, and maintain the current practice of not disclosing associated parameters (attributes, rates, or constraints). Because DataSwap shows

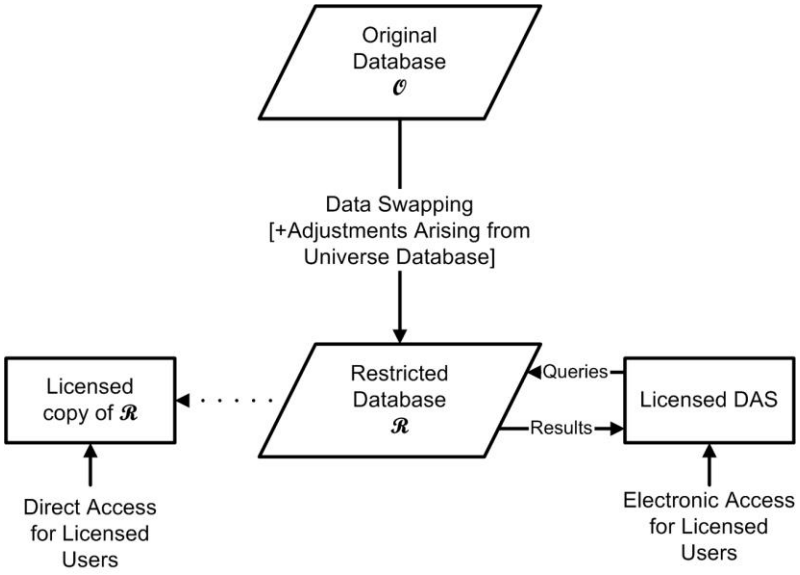
 ISSN 2047-3745

substantial replicate variability of data utility with other parameters held constant, the task force suggested that NCES run DataSwap multiple times on O to produce multiple candidates for R, and select a candidate with desirable data utility and disclosure risk characteristics.

The rationale for producing R from O is to ensure a small level of disclosure protection for all data records with minimal loss of data utility. Data swapping, assuming all records have positive probability of being swapped, is an appropriate means of statistical disclosure limitation (SDL) for accomplishing this, especially for categorical data. In particular, no one-dimensional distributions are altered and most (although data users do not know *which*) multidimensional distributions are not altered.

Access to Restricted Databases. Whether directly or by online DAS, access to restricted databases should be under license from NCES. Figure 2 illustrates the resulting structure.

Figure 2: Recommended structure for programs in which there is *no public database* and only licensed access to the data



Using a Licensed Online DAS to Access Restricted Databases. An online DAS accessible only to licensed users has two compelling strengths:

1. Such a data access system can be of unlimited statistical power, with full scripting capability and multiple user interfaces, including graphical interfaces. In fact, a licensed online DAS of unlimited power would obviate the need for physical transfer of data from NCES to licensees, eliminating security and monitoring issues. Of course, a DAS of "unlimited statistical power" might be prohibitively expensive and complex to create and maintain.
2. By recording and analyzing queries processed by the online DAS, NCES would have a window into usage of its data that does not currently exist, and which could inform the design and improve the quality of future data collections.

 ISSN 2047-3745

The task force acknowledged that a licensed online DAS poses issues of authentication and encryption, but that current technologies are adequate to deal with them.

Using a Public Online DAS to Access Restricted Databases. The task force noted that given current understanding of the disclosure risks associated with data access systems (Gomatam, Karr, Rieter, & Sanil, 2005) (Karr, Kohnen, Oganian, Reiter, & Sanil, 2006), an online DAS allowing public access to R would have to be severely limited in terms of allowable queries and responses to be deemed safe.

Limiting a public online DAS to allow access to restricted databases would require among other restrictions:

1. *Subsetting of the data.* This is an issue for individual queries; for instance, the mean income of a small number of subjects is more informative about individual incomes than the mean income for a large number of subjects. More subtly (Reiter, Organian, & Karr, 2008), it is also an issue of *query interaction*: by comparing the results of two queries on subsets of the data differing by one subject, information about that subject is revealed. Preventing the first problem is straightforward;[1] the second is not, since in particular it requires tracking the entire query history for the DAS.
2. *Transformations of variables.* As discussed in (Gomatam, Karr, Rieter, & Sanil, 2005), high-leverage transformations of variables entering regressions can reveal individual attribute values. In some ways, such transformations are simply an implicit way of subsetting the data. The query space of a DAS can forbid or severely limit transformations of variables, for instance, by allowing only standard transformations (square roots and logarithms) used to make data "more normally distributed."
3. *Interactions.* While less is known about risks associated with interactions than those arising from subsetting and transformations, there is a problem. Interactions of arbitrarily high order also, in effect, subset the data. The query space of a DAS can limit the order of interactions, although there is no clarity about how much restriction is "enough."

Current knowledge is not sufficient to guarantee safety. However, there are two classes of data access systems about which there may be enough known for NCES to proceed:

1. *Table servers* (Dobra, Fienbert, Karr, & Sanil, 2002). In this case, it is very likely that limiting the dimension (for example, to three or less) of tables provided is adequate to protect confidentiality.[2]
2. *Regression servers* (Gomatam, Karr, Rieter, & Sanil, 2005). In this case, of course, restriction of the output is mandatory; for instance, residuals cannot be released. How cautious to be in dealing with the subsetting/transformation/interaction issue is not obvious. However, it is possible that the user community for a public DAS running on R would be satisfied with extreme caution of the form "no subsetting,[3] no transformations other than square roots and logarithms, and no interactions."

If the query space of a public online DAS running on R is sufficiently restricted, the task force urged that NCES consider pre-computing the answers to all—or a large number of—queries, and having the DAS access these rather than R itself. Doing this reduces security issues arising from public access to a system that interacts directly with R. To illustrate, a table server that provides only tables of dimension 3 or lower from

[1] An alternative approach, called *differential privacy*, has been proposed in the computer science literature (Dwork 2007), in which noise is added to query results, and the noise level (variance) is higher the fewer records involved in the query.
[2] Even this is not certain, because methods from computational algebraic statistics can provide information, sometimes very precise, about individual entries in the full table.
[3] In the case of a database containing both numerical and categorical variables, subsetting only on the categorical variables and only when the associated "cell count" exceeds a threshold.

 ISSN 2047-3745

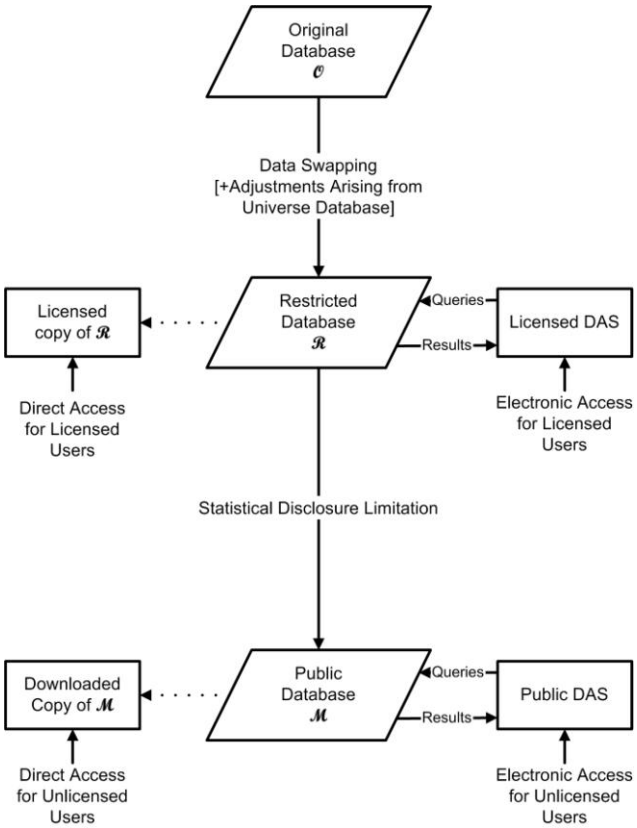
an R with 100 dimensions would need to pre-compute and store the answers to only approximately 1 million queries.

Creation of Public Databases. The task force suggested producing public databases, M, including weights (modified to avoid disclosure risk) whenever possible.

Transformation of R to M. The public database M should be produced from R only by means of (a) deletion of (sensitive or other) attributes; (b) category aggregation for categorical variables; and (c) coarsening for numerical variables. These SDL methods enforce consistency between the same analyses performed on both, in large part because each restricts the set of analyses that actually can be performed on both R and M. Consistent results are not only a quality control goal, any inconsistencies may also be informative about the SDL used to produce M from R, and therefore subvert the safety of M.

Explicit information about SDL strategies used to transform R to M as recommended above should be provided to users as it would pose no threat to confidentiality.

Figure 3: Recommended structure for programs *with* public databases



Given the commitment of NCES to licensing, a public database M can serve a client base very different from that of R. Users of M may be relatively unsophisticated or undemanding statistically. For instance, they may seek only tabular or other summaries rather than detailed statistical models; or, they may be students wanting to explore and understand "real data." The implication is that the statistical disclosure limitation used to produce M from R can be rather strong, ensuring that disclosure risk is negligible.

Releasing M without weights destroys its utility for any purpose. However, existing SDL theory and methodology have largely ignored the role of weights with respect to disclosure risk (de Wall & Willenborg, 1997). Clearly the values of weights themselves represent disclosure risks. For instance, weights may be informative about geographical detail that has been removed from M.

Access to Public Databases. The task force suggested that NCES provide online DAS access to public databases whenever possible. The structure in this case is shown in figure 3. DAS access to M is not a logical necessity, since anyone who wants to can download M. There are, however, two strong reasons for NCES to provide a DAS for public databases:

1. by recording and analyzing queries processed by the DAS, NCES would have a window into usage to its data that does not exist currently, and which could inform the design and improve the quality of future data collections.; and
2. service to clients, since some users of the DAS may not be able to perform the corresponding analyses, no matter how simple, on M.

There is, however, a consistency issue: if Q is a query that a DAS operating on M will respond to, then the result of Q applied to M must be identical with the result of Q applied to R.

Neither the query space nor the results provided by a public online DAS running on a public database M need be restricted in order to limit disclosure risk, since any user can perform the same analysis on M. Instead the inherent nature of M limits queries and results.

Learning from the uses of R **and** M **and their respective DASs.** All data access systems should be configured to collect as much information about database uses as possible. NCES should use such information to inform modification of existing data products and design of future ones in order to tailor the user interfaces of data access systems to user communities. For uses of public databases (M) in particular, the process of developing better knowledge could include polls of users as well as uses collected public DASs in order to better tailor the SDL methods used in the transformation of R to M.

Related NCES Activities

Prior to 2003 it was common NCES practice to produce both public-use and restricted-use versions of its datasets. Getting access to the restricted-use version required obtaining a license to assure the security of the data and compliance with applicable confidentiality laws.

At the beginning of 2003, NCES made a decision to make all datasets available in one place online and develop an online data access system (DAS) with a common look and feel to eliminate the need to learn multiple formats and access procedures. The existing (2002) data policy was that if the data underlying an online DAS was a restricted-use file (RUF), then no public-use file (PUF) could be released. Having only PUF data underlying an online DAS as well as having both a licensed RUF and PUF with no DAS access

 ISSN 2047-3745

was also an allowable configuration. (Is should be noted that at the time, some NCES data sets were being released (via download or CD-ROM) with very rudimentary data analysis software included; the acronym "DAS" was also used for this software. It should not be confused with the online Data Access System.)

NCES asked NISS to form a task force to help review the adequacy of its 2002 data confidentiality and licensing standards and procedures. This task force's findings and suggestions are summarized above.

As suggested by the task force, NCES continues to base all of its analyses and publications on RUFs that have been subjected to some basic statistical disclosure limitation procedures. NCES did adopt the task force's suggested modifications to the SDL procedures involving its DataSwap software.

An online DAS option is now available for many NCES data sets. NCES chose not to implement the task force's preference that the DAS be driven by RUFs, require a license to use, and have no restrictions on its analytic capabilities. To maximize access to its data through the online DAS, NCES chose to limit the DAS's capability to simple table building and simple linear regression (producing correlations without interaction terms or residual reporting) so that it could run from a RUF without requiring a license. The online DAS is structured so that almost all NCES complex survey datasets can use its functions. However, because analysis of assessments require specialized statistical procedures, a special purpose online DAS called the NAEP Data Explorer (NDE) has also been developed for access to data from the Main NAEP (National Assessment of Educational Progress), Long Term Trend NAEP, and the NAAL (National Assessment of Adult Literacy). Unlimited analysis of RUF data is only available for those obtaining a copy of a RUF under license.

Although most public-use files released prior to the DAS decision remain available, there are a few surveys that are continuing the practice of releasing both RUFs and PUFs (i.e., ECLS-K, NHES, FRSS, PIRLS, PISA, PSS, SSOCS, and ALS). However, in these cases, neither the RUFs nor the PUFs are accessible via an online DAS.

The NCES website also provides some table lookup capabilities for its universe data (CCD, IPEDS, and PSS) and some limited custom table construction and mapping capabilities. These universe administrative data sets are not collected under a pledge of confidentiality.

The best way to learn what data sets are available is to go a survey's main web page on the NCES site (usually found most easily by searching on the survey acronym or full title in the search box on any NCES web page. Once on the selected survey's page, there may be a direct link to "Data Analysis System (DAS)" in addition to a "Publications & Products" link. The former will lead to the NCES common DAS page where you may explore which data sets are accessible through the online DAS. The latter will bring up another list, among which is there will be a "Data Products" link. Clicking on that link will lead to a list of all available data sets produced by that survey and their status as "restricted-use" (requires licensing) or "public-use."

Task Force on Effect Sizes

Issue

The usefulness of reporting effect sizes, a concept widely used in educational psychology and other education research, in NCES publications

 ISSN 2047-3745

Charge
The principal goals of the NISS task force on effect sizes were to assess whether—and if so, how—NCES should report effect sizes in its publications (Karr A., NISS/NCES Effect Size Task Force: Final Report, 2008).

The following specific questions were to be addressed:

- For which results should NCES data collection programs report effect sizes?
- What are appropriate measures of effect sizes for particular results?
- In what way(s) could effect sizes be presented (including visualizations) and interpreted in NCES publications?

Membership
See Appendix 2 - Task Force Membership.

Task Force Findings and Suggestions
Overall. In general, the task force strongly supported reporting of effect sizes by NCES, but realized that there are instances in which doing so may be inappropriate, ineffective, or even impossible. At the same time, the task force acknowledged that the very dimensionless characteristic that makes effect sizes attractive for some purposes raises issues of interpretability that cannot be dismissed lightly. For instance, most people can understand and interpret a difference of $250, but not an associated effect size of 0.85. Researchers are looking into interpretability metrics (Bloom, Hill, Black, & Lipsey, 2008)

For Differences in Means. When underlying values lack physical interpretability[4], routinely report effect sizes for differences in means *instead of* the actual differences; when underlying values have strong physical interpretability, routinely report effect sizes for differences in means *in addition to* the actual differences.

- Use standardized mean differences as the effect size measures.
- Do not report effect sizes for differences in means if the associated (absolute) difference is not statistically significant or below a designated detection level.
- Tables containing effect sizes for differences in means should *not* also contain the actual differences in means.

For Differences in Category Proportions. When underlying values lack physical interpretability, routinely report effect sizes for categorical comparisons *instead of* the actual comparisons; when underlying values have strong physical interpretability, routinely report effect sizes for categorical comparisons *in addition* to the actual comparisons.

- The preferred measures are standardized differences in proportions.
- Do not report effect sizes for differences in proportions if the associated (absolute) difference is not statistically significant or below a designated detection level.
- Tables containing effect sizes for differences in proportions should *not* also contain the actual differences in proportions.

[4] An example is assessment scores, as compared, for instance, to dollar amounts or student enrollments.

Uncertainties in Effect Sizes. Identify circumstances under which reporting uncertainties associated with effect sizes improves quality and usability, and do so in such cases. Such circumstances depend on multiple factors, including diverse audiences and purposes for NCES publications, in ways that seem to preclude succinct summary.

The task force found that there is currently no consensus regarding how to convey uncertainties to literate but not statistically sophisticated individuals. Potential methods, which are largely identical in terms of mathematical content, include: confidence intervals; effect size ± uncertainty, where uncertainty, typically, is confidence interval half-width; effect size ± uncertainty/effect size (i.e., effect size ± *p*%); and, graphical methods (e.g., map effect uncertainty onto color, although this may conflict with existing report and website standards).

There are counterbalancing issues of information overload. The inclusion of uncertainties may decrease usability for those not interested in or not able to assimilate uncertainties.

Large Effect Sizes. Evaluate the feasibility of developing and implementing a sensible, consistent mechanism for calling attention to large effect sizes.

- Do not employ arbitrarily defined cutoffs or value-laden characterizations to define "large" effect sizes, as has been proposed in some papers in the literature (Cohen, 1988).
- "Large" should represent either exceeding a *touchstone* – a scientifically defined important difference, such as one year's progress in an assessment context – or an extreme value relative to a population of effect sizes, for example, among the *x* percent of effect sizes in a report, or in a collection of similar reports.

Scientifically based touchstones, when they can be identified, are preferable. Among the problematic aspects of "extreme relative to some population" are (1) defining that population, (2) that whether an effect size is large depends on how many other effect sizes are reported, and (3) that spurious large effects are possible. Other problems include: "Large" may be interpreted as "important," or even "causal"; and, "large" could have different interpretations in different contexts.

Related NCES Activities

Many of the comparisons (differences in means or proportions) reported by NCES are expressed in familiar and easily interpretable units, e.g., dollar amounts, enrollments, graduations, etc. This obviates the need for and takes a lot of publications out of consideration for using and reporting effect sizes.

The NCES longitudinal studies program has done exploratory research to define appropriate effect size statistics, develop useful touchstones to identify large effect sizes, and create constructs that enhance interpretation of effect sizes in useful contexts. Three areas of research interest are:

- Selecting a denominator for effect sizes of differences in means and proportions,
- constructing an interpretable frame of reference for magnitude: e.g., achievement improvement per grade, and
- using quartiles as anchor scores for meaningful comparisons.

The National Assessment of Educational Progress (NAEP) has done exploratory research to develop an effect size metric for NAEP. Four areas of research interest are:

 ISSN 2047-3745

- measuring effect size as a proportion of expected annual cognitive growth,
- linking NAEP and ECLS-K survey assessment correlations,
- metrics for within year studies, and
- metrics for within cohort analysis.

The Early Childhood Longitudinal Study – Kindergarten survey program has used a fixed cut-point for evaluating differences.

Task Force on Computerized Adaptive Testing (CAT)

Issue
The feasibility and potential value of using CAT for assessments in NCES longitudinal studies.

Charge
The principal goals of the task force were to consider the use of CAT in NCES longitudinal studies in general and in High School Longitudinal Survey (HSLS-09) in particular (Sedransk, 2008).

Membership
See Appendix 2 - Task Force Membership.

Task Force Findings and Suggestions
Adaptive Computer-Based Testing. CAT is a particular form of computer-based testing (CBT), which is now a mature, well-tested set of technologies. In contrast to "paper" testing, CAT offers advantages such as

- use of interactive items;
- opportunities to assess e-learning and e–learning processes;
- options to present information over time to evaluate the new information and reevaluate earlier information;
- opportunities to assess attentiveness and/or the incorporation of mechanisms to improve attentiveness; and
- security of information, both to minimize data loss and to immediately recapture lost or inconsistent information (rather than by means of recalls or repeated visits), with careful attention to computer security especially vis-à-vis information transfer.

CAT is distinguished by the adaptive selection of items for an individual test-taker based on previously collected external information and on responses to earlier items.

CAT is appropriate for (1) broad assessments requiring comparable precision over a fairly wide range of performance, and (2) assessments being widely used for analysis, including analyses of data subsets or of population subgroups. Longitudinal studies typically have both these attributes. Another principal advantage of CAT is efficiency in terms of time limitations for student assessment (or an equivalent improvement in scoring precision). Various types of adaptive designs (i.e., two-stage adaptive designs, item adaptive designs, testlet adaptive designs, or variants of these) carry different relative advantages. The choice in a

 ISSN 2047-3745

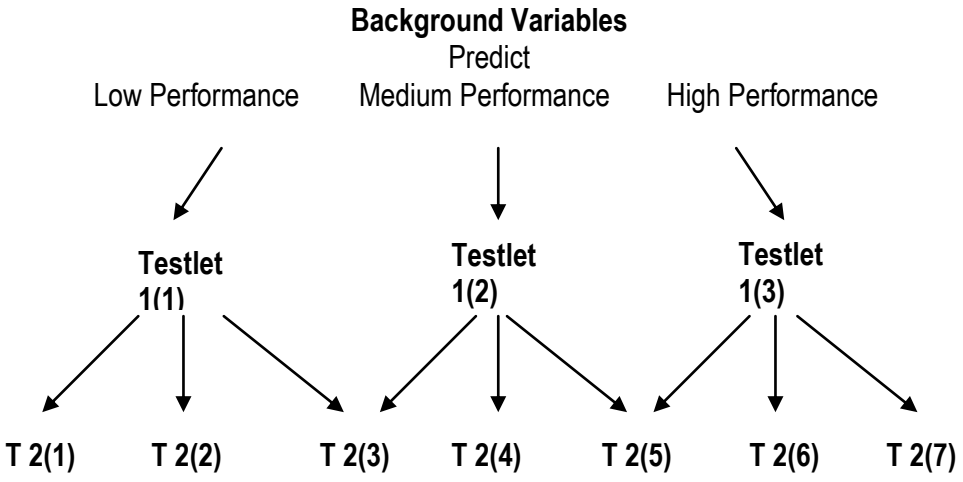
particular instance depends on the specific assessment goals, the structure of the test, and the time available for administering the test.

A major advantage of CAT is that adaptation can occur with respect to previous responses as well as to information from external sources (student background and history, previous assessment results, etc.). By integrating this external information into the design, CAT may be able to provide acceptable precision even in the ever-shorter assessment times, for example, by sharpening the "zero-th" stage process to identify with approximate accuracy the student's performance level to initiate testlets closer to the actual (true) performance level.

CAT requires an early commitment to an adaptive form. Item construction and item calibration in computer-administered form must precede or proceed in parallel with formulation of the adaptive structure and then with the algorithm development. Software development also requires early decisions about platform(s), security requirements, ancillary features (such as monitoring attentiveness), data transmission verifications, and the ultimate database structure. The benefit of this early preparation is streamlining the compilation of the database by minimizing difficulties in data transmission, data editing, and database creation.

CBT, including CAT, can be conducted on-site or remotely via the World Wide Web. The use of secure Internet data *transmission* is usually crucial and feasible. In contrast, web-based test *administration* (remote or on-site) is seriously flawed, carrying with it the potential of many costly difficulties that have not yet been resolved. These range from coordination of administration to security of items and responses, authentication of test-takers' identities, adequacy/equivalence of multiple platforms, as well as disruptions to Internet performance and other problems outside the control of the test administration staff.

Example: The 2.5-stage Design Model of CAT. The "2.5-stage test design, discussed in some detail by the task force, exploits the idea that student background data (or "covariates") **Y** might serve as the basis for selecting a routing test, or the first testlet, in a conventional two-stage adaptive test (Lord, 1980). A graphic depiction of such an adaptive test is shown below.



 ISSN 2047-3745

In this arbitrary depiction, the student background data (or "covariates") **Y** are used to "route" students into one of three first-stage testlets [1(1), 1(2), or 1(3)] which are designed for students of relatively low, medium, or relatively high proficiency, respectively. The responses of the students on the first-stage testlet, $\mathbf{X}_{m1}$, possibly in concert with **Y**, are then used to route the student to one of a more finely graded set of second-stage testlets. In the graphic above, there are seven second-stage testlets, ranging from very low to very high in difficulty.

Unlike a conventional two-stage test, this design has two points of adaptation: The first is between the collection of the background data **Y** and the second is between the first and second stages of the test itself. It is not, however, strictly speaking, a three-stage test (although it has as many points of adaptation as a three-stage test). Hence, a name for this might be a "two-and-a-half stage test" or a "two-plus stage test."

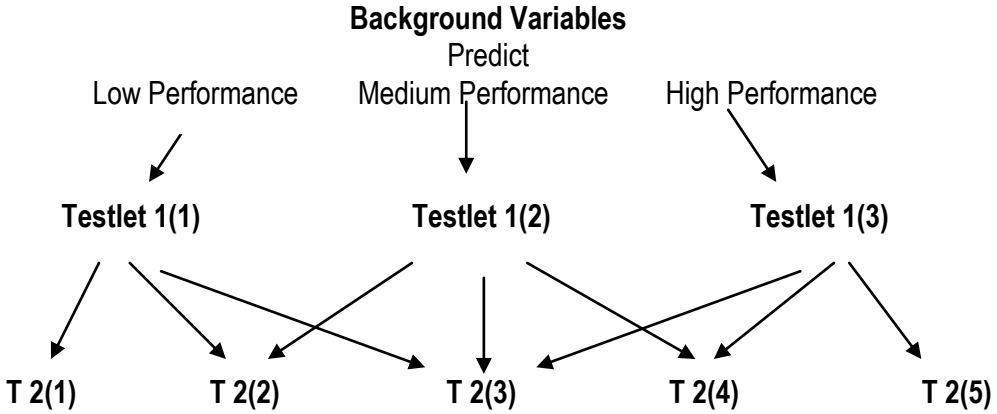
Two points of adaptation are far superior to the one that comes from the use of a two-stage test alone, because the second adaptation can "correct" for routing errors that may arise if there are only two stages and one adaptive choice.

Scores on this "two-and-a-half stage test" or a "two-plus stage test" may be computed with any number of approaches based on item response theory.[5] Scores may be based solely on the item response data **X** or on the combination of **X** and **Y**. The choice between those options may depend on the use of the scores.

The graphic below shows a more realistic two-plus stage test design, in which there are only five second-stage testlets, ranging from easy to difficult. Note that, comparing this more realistic design to that above, there is more overlap between second-stage testlets that can be reached from each pair of first-stage testlets.

In the construction of multistage tests, it is neither practical nor desirable to construct testlets that measure only a very narrow range of proficiency, nor to route examinees with extreme precision based on short testlets. The result is that, in practice, the proficiency distributions of students routed to each of the three first-stage testlets overlap substantially; so, at the second stage of testing, those students are routed to overlapping testlets.

[5] See (Thissen & Wainer, 2001), chapters 3, 4, 7, and 8 for a summary of alternatives.



Adaptive Testing for Science, Technology, Engineering, and Math (STEM) Constructs and Facets. Some aspects of educational assessment are largely independent of the mechanism of administration, be they CAT or other modes. These independent aspects include determining which constructs and facets to measure. In general, choice of construct aspects to measure can be justified in terms of

1. providing the largest amount of *additional* information, given information from other sources;
2. reflecting specific covariates of general importance and covariates that differ among subpopulations of particular interest;
3. directly relating to primary inferences to be drawn both for cross-sectional and for longitudinal analysis objectives; and
4. having educational relevance, both in the source for the construct and in the potential for actions to be taken based on the assessment results.

Some constructs seem to be uniquely or distinctly more easily measurable via CBT or CAT, such as mastery of processes involved in e-learning, of complex reasoning processes involving the evaluation of the value and validity of individual pieces of information, and of complex tasks for which the subtask sequence depends on the degree of complexity the student is able to recognize.

Any assessment in STEM settings is complicated by lack of a universal science construct. Science is not monolithic, nor are reasoning skills and interest levels fully shared across scientific areas. It has yet to be demonstrated that context-free assessment of scientific skills, especially the capability to reason with scientific information, can be accomplished. Breadth across sciences and depth in a particular area of science are both important.

Elucidating the interplay of any of these with a host of covariates (such as family structure, socioeconomic status, influence by family members, friends, teachers, and/or the popular media) seems to demand CAT, especially given the longitudinal nature of many of the research and policy questions.

 ISSN 2047-3745

Using CAT, a hybrid design in the STEM setting that addresses the issues just discussed, is possible and feasible. One solution is to use a multiscience single context, for example, an environmental problem, for several testlets. The first of these can be more general and presented to all students; subsequent testlets can be individualized to be specifically in each student's designated (preferred) domain.

Mathematics is essentially linear through the algebra-to-calculus level[6] college track for the algebra-geometry-trig-calculus college track sequence in the United States, although this is not the case for other tracks or for other (e.g., foreign) curricula. As a result, mathematics poses fewer difficulties than science in a high school setting.

Use of CAT for STEM raises serious but manageable issues of item sources and assessment design. Concerning the former, existing item pool—for example, the Programme for International Student Assessment (PISA), the National Assessment of Educational Progress (NAEP), the National Educational Longitudinal Study (NELS), and others—are extensive and well calibrated. With careful selection of items, these can serve as a valuable, but not sole, source of suitable items, noting that items with multiple plausible distractors are required for partial scoring. There is, however, an important caveat: *Item calibration for a traditional paper-pencil multiple-choice test is not automatically sufficient to calibrate the same items for inclusion in computer-adaptive tests, although it is often possible to get close enough to allow equating.* For constructs that can be assessed uniquely by computer-based or computer-adaptive testing, new items need to be constructed, with emphasis on responses that can utilize a scoring model that gives partial credit.

In terms of assessment design, evidence-centered design creates a structural approach to the design, implementation, and delivery of assessments. These principles form the basis for a "best practices" approach that is fully applicable to longitudinal studies and that can directly incorporate computer-adaptive test structures.

Implications for HSLS-09. The High School Longitudinal Study **(**HSLS-09) meets all the principal criteria for use of computer-adaptive testing. For students, HSLS-09 is a "low stakes" test: there is little incentive to cheat, but there is a correspondingly non-negligible likelihood of inattentiveness. The test administration time for HSLS-09 is severely limited. The two planned assessments (early 9th and late 11th grades) can be done with a single platform and supporting software. Extensive auxiliary information is to be incorporated into the database, which leverages the strengths of CAT by allowing pre-categorization of each student's specific scientific strength (preferred domain) and of the most appropriate level for entry into a CAT.

Task Force Conclusions

- CAT is *feasible*, because of technological advances, as well as students' facility with computers.
- CAT is *desirable*, reflecting the evolution of education mechanisms, practices, and goals.
- CAT is *effective*, often uniquely so, in the face of time limitations and other practical constraints.
- HSLS-09 is an *appropriate entry point* to CAT for NCES, because of its longitudinal nature and because its assessments are not used for evaluative purposes.
- HSLS-09 is an *excellent opportunity* to take advantage of the strengths and flexibility of CAT to integrate administrative records, prior test performance results, and interest information into the adaptive testing algorithm.

[6] AP statistics may be the only numerically significant departure from the algebra-geometry-trigonometry-calculus path.

Related NCES Activities

- Subsequent to the NISS report, NCES changed its planned 2010 assessment component of HSLS-09 from paper and pencil adaptive to CAT.
- NELS:88 and ELS:02 both used paper and pencil adaptive testing.
- Both before and after receiving the report, NAEP has run tests of CBT. In 2009 they fielded a national sample of the science assessment using CBT.
- ECLS-K was paper and pencil adaptive since 1998 and plans are to do the same for the next ECLS-K beginning in 2011.
- Both NELS and ELS were paper and pencil adaptive.

Related Task Force Member Spin-off Activities

- Dave Thissen was one of the speakers at the SAMSI Summer Program on Psychometrics, (SAMSI, 2009)
- Jonna Kulikowich and Nell Sedransk are collaborating with researchers at the University of Connecticut on metrics for online reading comprehension (including computer based testing although not necessarily adaptive).
- Jonna Kulikowich is collaborating with Elizabeth Stage on current projects.
- Kentaro Yamamoto has included others from the task force in a major computer-adaptive /electronic measurement project.
- Henry Braun, Kentaro Yamamoto, and Elizabeth Stage are now collaborators.
- Kim Gattis is working with Jonna Kulikowich on another large research project.

References

1. Bloom, H., Hill, C., Black, A., & Lipsey, M. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educaitonal interventions. *Journal of Research on educational Effectiveness* , 289-238.
2. Cohen, J. (1988). *Statistical Power Analysis for the Behaviorial Sciences.* New York City: Academic Press.
3. de Wall, A., & Willenborg, L. (1997). Statistical disclosure control and sampling weights. *Journal of Official Statistics* , 417-434.
4. Dobra, A., Fienbert, S., Karr, A., & Sanil, A. (2002). Software systems for tabular data releases. *International Journal of Uncertainty, Fuzziness, and Knowledge Based Systems* , 10(5) 529-544.
5. Gomatam, S., Karr, A., Rieter, J., & Sanil, A. (2005). Data dissemination and idsclosure limitation in a world without microdata: A risk-utility framework for remote access and analysis servers. *Statistical Science* , 163-177.
6. Karr, A. (2009). *NCES/NISS Task Force on Nonresponse Bias: Final Report.* Research Triangle Park, NC: National Institute of Statistical Sciences.
7. Karr, A. (2008). *NISS/NCES Data Confidentiality Task Force: Final Report.* Research Triangle Park, NC: National Institute of Statistical Sciences.

 ISSN 2047-3745

8. Karr, A. (2008). *NISS/NCES Effect Size Task Force: Final Report.* Research Triangle Park, NC: National Institute of Statistical Sciences.
9. Karr, A. (2009). *NISS/NCES Task Force on Configuration and Data Integration for Longitudinal Studies: Draft Final Report.* Research Triangle Park, NC: National Institute of Statistical Sciences.
10. Karr, A. (2009). *NISS/NESSI Task force on Full Population Estimates for NAEP: Final Report.* Research Triangle Park, NC: National Institute of Statistical Sciences.
11. Karr, A., Kohnen, C., Oganian, A., Reiter, J., & Sanil, A. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician* , 224-232.
12. Lord, F. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.
13. NCES. (2005). *User's Guide to Developing Student Interest Surveys Under Title IX.* Retrieved August 31, 2009, from NCES: http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005173
14. NCES/NISS Task Force on Non-Response Bias.
15. NISS. (2004). *Project Profile: Education Statistics*. Retrieved August 31, 2009, from NISS: http://niss.org/research/project-profile-education-statistics
16. NISS. (n.d.). *Projections of Education Statistics to 2017*. Retrieved August 31, 2009, from NCES: http://nces.ed.gov/programs/projections/projections2017/
17. NISS. (2002). *Review of NCES Statistical Standards*. Retrieved August 31, 2009, from NISS Project Profile: Education Statistics: http://niss.org/research/project-profile-education-statistics
18. NISS/ESSI Task Force. (2005). *NISS/ESSI Task Force on Graduation, Completion, and Dropout Indicators.* Retrieved August 31, 2009, from NCES: http://nces.ed.gov/pubs2005/2005105.pdf
19. Piesse, A., & Rust, K. (2003). *U.S. 2001 PIRLS Nonresponse Bias* . Retrieved August 31, 2009, from NCES: http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=200321
20. Reiter, J., Organian, A., & Karr, A. (2008). Verification servers: enabling analysts to assess the quality of inferences from public use data. *Manuscript in preparation* .
21. SAMSI. (2009). *Previous SAMSI Workshops and Short Courses*. Retrieved August 31, 2009, from SAMSI: http://www.samsi.info/workshops/prevsamsiworkshops.shtml
22. Seastrom, M. (2009). Tracing Survey Respondents Without SSNs. *Proceedings of the 2009 Joint Statistical Meetings.* Washington, DC: American Statistical Association.
23. Sedransk, N. (2008). *NISS/NCES Task Force on Computer-Adaptive Testing: Final Report.* Research Triangle Park, NC: National Institute of Statistical Sciences.
24. Statistical Standards Program, NCES. (2003). *NCES Statisitcal Standards Program.* Retrieved August 31, 2009, from National Center for Education Statistics: http://nces.ed.gov/StatProg/Standards.asp
25. Thissen, D., & Wainer, H. (2001). *Test Scoring.* Mahwah, NJ: Lawrence Erlbaum Associates.

Appendix 1 – Other NISS Work for NCES

Complete:

- Issues around multiple comparisons (unpublished work, circa mid-1990's);
- Calculating nonresponse rates and consequences of the NELS nonresponse bias analysis (unpublished NISS report to NCES on response rates across all NCES surveys, 2008);
- Review of NCES draft 2002 statistical standards (NISS task force, 2002), resulting in current NCES standards (Statistical Standards Program, NCES);
- High school graduation, completion, and dropout indicators (NISS/ESSI Task Force, 2005);
- NCES participation in international assessments (NISS task force, 2004) (Piesse & Rust, 2003);
- Title IX issues in post-secondary institutions (NCES, 2005);
- Nonresponse bias analysis (Karr A., NCES/NISS Task Force on Nonresponse Bias: Final Report, 2009); and
- Full population estimates (Karr A., NISS/NESSI Task force on Full Population Estimates for NAEP: Final Report, 2009).

In process:

- Coordinating future data collections to better represent the pre-K to post-secondary education pathway and adult education (Karr A., NISS/NCES Task Force on Configuration and Data Integration for Longitudinal Studies: Draft Final Report, 2009),
- Maps and graphs in NCES data dissemination (ongoing project),
- Projection of Education Statistics (ongoing project),
- Postsecondary access and choice (ongoing project),
- Tracking without Social Security numbers (Seastrom, 2009), and
- Example implementation of nonresponse bias analysis recommendations (ongoing project).

 ISSN 2047-3745

Appendix 2 – Task Force Membership

Members of the Data Confidentiality Task Force
Alan Karr, NISS (chair)
George Duncan, Carnegie Mellon University
Stephen Fienberg, Carnegie Mellon University
Bobby Franklin, Louisiana Department of Education
Gerald Gates, Census Bureau (now, private consultant)
Jerome Reiter, Duke University
Lynne Stokes, Southern Methodist University
Rebecca Wright, New Jersey Institute of Technology (now, Rutgers University)
Anna Oganian of NISS provided postdoctoral support.

Members of the Effect Size Task Force
Alan Karr, NISS (chair)
Mark Lipsey, Vanderbilt University
Stephen Olejnik, University of Georgia
Ingram Olkin, Stanford University
Bruce Spencer, Northwestern University
Bruce Thompson, Texas A&M University
Leland Wilkinson, Systat
George Luta of NISS provided postdoctoral support.

Members of the Computerized Adaptive Testing Task Force
Nell Sedransk, NISS (Chair)
Henry Braun, Professor, Boston College
Kim Gattis, NAEP ESSI, NAEP Assessment Development and Quality Assurance
Jonna Kulikowich, Pennsylvania State University
Robert Mislevy, University of Maryland
Elizabeth Stage, University of California at Berkeley
David Thissen, University of North Carolina at Chapel Hill
Howard Wainer, National Board of Medical Examiners
David J. Weiss, University of Minnesota
Kentaro Yamamoto, Educational Testing Service

 ISSN 2047-3745