NISS

NISS/NESSI Task Force on Full Population Estimates for NAEP

Final Report
Revised: July 22, 2009

Technical Report Number 172
July 2009

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

NISS
National Institute of Statistical Sciences
PO Box 14006, Research Triangle Park, NC 27709-4006
Tel: 919.685.9300 FAX: 919-685-9310
www.niss.org

NISS/NESSI Task Force on Full Population Estimates for NAEP

Final Report
Revised: July 22, 2009

0 Framing Statement

As understood by the task force, the goal of NAEP is to provide high-quality indicators of performance for well-defined populations of students enrolled in selected grades of U.S. schools.

Under current NAEP protocols, some students with disabilities (SD) and some English language learners (ELL) may be excluded from assessment. Inclusion rates differ across states. The task force believes that as a result the goal of NAEP is difficult to meet under current protocols, and they will become increasingly difficult to meet in the future. The task force further believes that NCES must ultimately choose between two alternatives:

1. Adjust reported NAEP findings to include estimates of the performance of SD and ELL students who were not tested, but reasonably could have been.

2. Redefine the population that NAEP claims to cover so that it does not include some SD and ELL students.

This report contains the task force's recommendations regarding this choice and related issues.

The task force strongly supports NCES' continuing to take a proactive approach to the problem of variable inclusion practices. Otherwise, important comparisons (state-to-state, year-to-year and subgroup-to-subgroup) may be distorted by differences and changes in inclusion practices.

1 Task Force Membership and Charge

The task force was convened by the National Institute of Statistical Sciences (NISS) under the auspices of the NAEP Education Statistics Services Institute (NESSI). Members are Robert Groves (University of Michigan), Robert Hauser (University of Wisconsin), Andrew Ho (University of Iowa), Lyle Jones (University of North Carolina

at Chapel Hill), Alan Karr (NISS and University of North Carolina at Chapel Hill, chair), Shelley Loving-Ryder (Virginia Department of Education), and Martha Thurlow (University of Minnesota).

The task force was charged to "recommend to NCES whether and how NAEP should construct and report full population estimates (FPEs)," by addressing such questions as:

- Are FPEs a valid scientific construct? Are students with disabilities (SD) and English language learners (ELL) conceptually different issues for FPEs?
- Should FPEs be reported at all, and at what levels of resolution? How should associated uncertainties be reported?
- Is sound statistical methodology available to calculate FPEs, which can also address issues such as adjustment of weights?
- Should FPEs be reported in addition to or instead of estimates based only on students who actually took the test?
- If FPEs are reported, what interpretations or warnings should accompany them?
- If FPEs are reported in addition to current estimates, how should the relationship between them be portrayed? In particular, would one be presented as subsidiary to the other? How should inconsistencies be presented and interpreted? What are the policy implications?
- Should FPEs be used at all or only some levels of resolution (geographical, sub-populations, ...)?

2 Background

The task force endorses efforts by NCES to conduct NAEP in ways that:

1. Increase inclusion, for example, by providing additional accommodations to students.
2. Make inclusion practices (important point: *not* inclusion rates[1]) more uniform across states and other reporting units.
3. Work through state NAEP coordinators to increase awareness of inclusion-related issues and their importance.

These efforts are important regardless of which of the two paths in section 0 is ultimately followed. However, the task force believes that it is not likely that these efforts will obviate the need to make the decision described there. Indeed, with increases in immigration and integration of students with disabilities into regular school programs, issues associated with inclusion may become worse. For instance, increased immigration increases the pool of potential ELL-*identified*, and therefore also ELL-*excluded* students. Similarly, an increase in inclusions may exacerbate difficulties in modeling of score distributions for excluded students: imputation will become more problematic because

[1] Rates are relative to SD- and ELL-identified populations.

characteristics of included and excluded students will differ more.

3 Major Recommendations

The task force's major recommendations are rooted in its strong belief that, of the two alternatives posed in section 0, **NCES should choose alternative 1—construction and publication of adjusted estimates**.

We further recommend that these adjusted estimates be named *expanded population estimates (EPEs).* The task force believes that NAEP estimates are meant to describe the population of students who reasonably could have been assessed, even if only some of these students are actually assessed. The term "full population estimate" is a misleading description of this population, especially to non-specialists, because it seems to imply the entire population of students. "Expanded population estimate" is more accurate; however, this term does not explain clearly what "expanded" is relative to. (The most precise term, "exclusion-adjusted estimates," seems simply to carry "too much baggage.")

The task force finds that alternative 1 is more consistent with the goal of "provid[ing] high-quality indicators of performance for well-defined populations of students enrolled in selected grades of U.S. schools." By contrast, adoption of alternative 2 would require a clear, explicit, and defensible definition of the reduced population that NAEP would then purport to assess, as well as an operational design consistent with that definition. The task force believes that it is not appropriate to define the target population of NAEP as those students who happen to have been judged capable of being assessed on a given occasion. Nor does it seem defensible to redefine the population that NAEP describes in a way that excludes identifiable categories such as SD and ELL students. Such a redefinition would replace the problem of differing inclusion practices by that of differing identification practices. Possibly, NAEP could define the population using NCLB conventions, but this raises other issues.[2]

The task force further recommends that **NCES set as its goal to report EPEs as the primary (or only) measure of NAEP performance.** Doing so requires an implementation of EPEs that is sound from both policy/decision and statistical

[2] NCLB recognizes that there are students with significant cognitive disabilities who are held to different achievement standards and not able to take the regular state assessments In some states, these students are identified on the basis of needing a modified curriculum, and those students with significant cognitive disabilities are assessed using alternate state assessments constructed to measure performance against alternate achievement standards. For adequate yearly progress (AYP) calculations at the state or local education authority (LEA) level, the number of such designated as proficient or advanced cannot exceed 1% of all students assessed at that level. NCLB further recognizes that there are additional students with more moderate disabilities who are learning grade level content but who will not achieve proficiency in the same time frame as their non-disabled peers. States are permitted to assess this latter class of students with alternate assessments measuring modified achievement standards; the number of designated as proficient cannot exceed 2% of all students assessed at that level. Although it may be appropriate to specify that students in the alternate assessment based on alternate achievement standards would not be eligible to be selected for NAEP, it is probably not appropriate to do the same for students in the alternate assessment based on modified achievement standards.

perspectives. The task force finds that methods used currently to calculate what are now termed FPEs are subject to criticisms raised regarding current NAEP estimates, as well as additional criticisms specific to FPEs, including the specific way in which they are constructed. Achievement of this goal requires that NCES resolve a set of recognized statistical issues, which are discussed in section 4.

The selection of "primary (or only)" rests ultimately with NCES. NCES' pursuing alternative 1 designates EPEs as the *more important* estimates for scientific and policy purposes.) The task force believes that reporting both EPEs and current estimates to some extent contradicts this position, and raises the risk, if the two are inconsistent, of creating confusion and skepticism about NAEP. However, the task force acknowledges that there may be factors of which it is not aware that argue for retaining current estimates, but relegating them to a position similar to the current position of FPEs. Once EPEs become the only (primary) set of estimates, the "expanded population" qualifier should be dropped.

The task force recommends that in the short run, **NCES continue to publish EPEs with the same degree of prominence as FPEs are currently published**—on the NCES web site rather than in printed reports, and with some informed effort necessary to locate them. As discussed in section 4, and notwithstanding the issues raised there, the task force finds that methods used to calculate FPEs are sufficiently sound that there is no identified need for drastic modification. NCES may wish, however, to strengthen disclaimers that EPEs remain under development. It may also wish, in tables containing actual EPEs, to label them as "Trial EPEs" in order to highlight the possibility that there may be future changes either to the data from which they are constructed or to the methodology used to construct them. NCES may also wish to provide a more detailed discussion of the nature and purposes of EPEs.

Finally, the task force recommends that **NCES move as rapidly as possible to conduct studies that sharpen understanding of the statistical issues described in section 4**.

4 Issues Regarding Calculation of EPEs

The task force believes that NCES can pursue alternative 1 if it is able to employ a scientifically and statistically sound, defensible methodology for estimation of plausible values for students who are (1) *Selected* for NAEP; (2) *Identified* as SD or ELL; and (3) *Excluded* from taking NAEP.

Currently, there exist two specific, and very similar, model-based approaches to imputation[3] of plausible values for SD/ELL students, which are described in McLaughlin (2005) and Braun, et al. (2006). We term these the core and extended methods, respectively. Their common general form is summarized in appendix A. Both create estimated plausible values by simulation, using

[3] We employ the term "imputation" even though doing so may go beyond ordinary usage. In a narrow sense, imputation is estimation of values that are known to exist, such as income, but are missing from a database. An assertion that NAEP plausible values exist but are simply not known for all excluded students is more tenuous. See also **Scope of Imputation** below in this section.

- A regression model to estimate mean plausible values;
- A separate variance estimation model to estimate variances.

In what follows, we refer to this shared structure as the current modeling framework (CMF). As noted in Wise (2006) and elsewhere, underlying the CMF is the assumption that *excluded students' performance is related to "available background information" in the same way as that of identified but tested students*. As discussed below, the task force is not entirely convinced that this assumption holds now, or that it will hold in the future.

Ultimately, methodology underlying EPEs will either be essentially within, although possibly extending or refining, the CMF, or else it will differ *qualitatively* from the CMF. However, the line between these two is blurred, and construing the choice as "retain or replace the CMF" is an oversimplification.

The task force recommends **that NCES view remaining within the CMF as the preferred course**, subject to resolution of issues discussed below in this section. The CMF is a strong foundation on which to build. There seem to be no major flaws that render it unusable or indefensible.

The task force further recommends that **NCES focus on the extended methodology for imputing plausible values**. The principal difference between the core and extended methods is inclusion of a school-level achievement[4] variable in the extended model that is not present in the core model. (Other differences—see also Wise (2006)—are that the extended methods fits five separate regression models for five separate plausible values and that the extended variance estimation model includes sampling variance. There may also be slightly different ways of handling missing predictors.) The extended model is more general, and several studies report that variable selection procedures retain the school-level achievement variable. As described in Wise (2006), while performance of the two methods in a setting of simulated exclusions is not dramatically different, the extended method seems to have smaller bias.

The task force has identified several specific concerns that it believes NCES should resolve, and will benefit from resolving, before committing to the CMF as the basis for constructing EPEs.[5] It is possible that sufficient information exists to resolve some of them now, and that the task force was simply unaware of such information. Theses issues do require research (and therefore, time and financial resources) to resolve. NCES' strongly principled commitment to sound science that informs sound policy supports undertaking the research, rather than exposing NCES to avoidable and potentially significant risk.

Validation. The CMF appears to have undergone rather limited validation, using simulation-based approaches described in Wise (2006). These approaches simulate

[4] On state assessments.

[5] The task force believes that it is also in NCES' interest to resolve these issues to its own satisfaction before committing publicly to a specific EPE methodology, rather than have them be raised externally.

exclusions, and then compare imputed to actual plausible values. They are therefore sensitive to the exclusion procedure. NCES would be in a stronger position if it were to deploy a modeling framework that has been validated thoroughly. One approach would be to administer NAEP to some students who were excluded, and to compare actual with imputed plausible values.[6] An alternative approach would be cross-validation, that is, to withhold some data points from the model fitting, which in effect turns them into exclusions, and to assess the model by comparing imputed values with true ones. Cross-validation would also shed light on the **Scope of Imputation** question.

Scope of Imputation. For which identified but excluded cases should plausible values be imputed? The CMF imputes for *all identified but excluded* cases, even though some of these differ drastically from *identified but included* cases. Wise (2004; 2006) shows that the extended and core models are both sensitive to the assumption that "excluded students have achievement levels identical to those of tested students *who are similar with respect to available background information*" (Wise, 2006; italics in original). Moreover, to impute for all excluded cases is inconsistent with "not tested, but reasonably could have been" proviso of alternative 1 in section 0.

Using classification algorithms, propensity scores or another method, it might make sense to limit imputation to cases for which the evidence for imputation is strong.[7] The impact would be a more defensible, albeit also more complex, methodology. Adjustments to weights might be necessary in order to accommodate this. Importantly, the issue is addressable in extant data sets, using simulation.

Missing Predictors. The task force is not convinced that there is adequate understanding of the impact of missing predictors in the CMF, or of other data quality issues such as faulty entries in ELL and SD questionnaires. Full resolution of the issue appears extremely challenging, but simulation could be used to gain some sense of the scale of the effects.

State-to-State Differences. Given that a driver for use of EPEs is state-to-state[8] differences in inclusion (and/or identification) *practices*, how are these differences best modeled? The CMF uses state-dependent variable centerings and model intercepts, which even if it does not maximize modeling precision, may not introduce major errors. The effects are simply not understood. Alternatives that can be explored with existing data include Bayesian models (hierarchical, random effects, …). Even the finding that increased model complexity yields no greater understanding would be valuable to NCES.

Using More Attributes. The task force finds especially promising the possibility of using additional student attributes as predictors in models employed for imputation, especially given the rapidity with which state-level student databases are being instituted. Approaches discussed included a pilot study using a state-level database (Florida was

[6] Stancavage, et al. (2007) reports a small-scale effort of this nature.
[7] Without going into specifics, the strength of evidence might be measured in terms of how "far away" the student lies from the data on which the models used to perform imputation are fit.
[8] Or other reporting unit.

mentioned as one candidate) containing additional attributes such as performance on a state assessment or higher-quality versions of the SD/ELL questionnaire attributes. See also discussion in the appendix of a screener exam.

Standard Errors. Some analyses seem to show that CMF-imputed plausible values have lower estimated standard errors than actual ones. To the task force and some others, this seems paradoxical. However, the situation is complex and subtle, and the effects appear not to be profound. Multiple imputation would assist in characterizing imputation variance, although this does not seem to be a sensible way of ultimately producing imputed plausible values. In any case, a more thorough understanding would be valuable before the CMF were be used in a production setting.

Weights. The task force is not certain what weights are used in model fitting within the CMF, or what weights are used for imputed plausible values when EPEs are calculated. Specifically, are these adjusted for absentees and refusals?[9] In part, this may be an operational issue resolvable through careful documentation. It is also not clear whether weights can or should be used within the imputation process.

Differences between Included and Excluded Students. The task force believes that better understanding of differences, within the population of SD- or ELL-identified cases, between included and excluded students will improve both efficacy and understanding of EPEs. There is also an important practical reason to address this question: the extent to which the student attributes in the CMF are predictive of exclusion is suggestive of possible weaknesses in this (or any other) framework. The problem is that the attributes being used to extrapolate from one population to the other may be the very ones that differentiate between them Including a propensity-of-exclusion score as a predictor might help address this issue.

Using Data from More Students. Can the CMF be improved by using models fit to data from *all* students taking NAEP, rather than only those who are identified as SD or ELL? While appealing conceptually, this path is problematic because the primary student-level attributes used as predictors in the models come from questionnaires administered only to students identified as ELL or SD.

Differences between SD and ELL. The task force is not certain that SD and ELL are sufficiently similar phenomena, either conceptually or operationally, that the same modeling framework is appropriate for both. As more means of accommodation are employed for SD students and as the population of ELL students increases, differences may become more pronounced. The evolution of the core approach, which initially employed one model for (SD only and SD+ELL) and one for (ELL) only, but later used one model (ELL only and ELL+SD) and one for (SD only), seems to confirm the lack of scientific understanding. Possibly, this is an issue requiring monitoring by NCES rather than immediate action.

[9] Although it is a peripheral issue, in calculating EPEs currently, NCES does not treat absentees and refusals in the same way it treats exclusions. The former are handled by adjusting weights of those not absent or refusing, and the latter by imputation. It may be important to have a good justification for this.

Among these issues, the task force views **Validation**, **Scope of Imputation**, **State-to-State Differences**, **Using More Attributes** and **Differences between Included and Excluded Students** as having the highest priority. **Standard Errors** and **Weights** are crucial, but are of a somewhat different nature and require fewer resources.

The task force is not in a position to make recommendations about the nature of new modeling framework if NCES were to determine that one is necessary. Above all, development a new modeling (or drastically altered) framework—which would happen only if the CMF were found grossly inadequate—would pose daunting issues of cost, time and uncertainty of success. However, the task force notes that:

- Conceptually attractive alternative modeling paths do exist. One of the most intriguing is *to impute item responses for excluded students*, from which plausible values would be derived using existing methods. To do this is appealing scientifically because the modeling is at a more basic level and also captures the full complexity of NAEP instruments. Clearly, however, such models would be intricate and computationally demanding, and ultimately there might be a paucity of data.
- A new framework may be driven principally by new sources of data rather than modeling *per se*. The task force finds the possibility of a NAEP screener exam[10] for ELL especially promising, and urges that NCES continue its consideration of such an exam, as well as consideration of a screener for students with disabilities.
- Modeling the identification and inclusion processes themselves may be valuable. The task force observes that even though two processes are taking place (identification and inclusion), almost all attention seems to have focused on the latter. Differences in identification practices may be as important as those in inclusion practices.

[10] By this, the task force means an instrument administered within NAEP that would provide information to support identification and inclusion decisions.

References

Braun, H., Zhang, J., and Vezzu, S. (2006). Evaluating the Effectiveness of a Full-Population Estimation Method. Unpublished paper, Educational Testing Service.

McLaughlin D. (2005). Properties of NAEP Full Population Estimates. Unpublished report, American Institutes for Research.

Stancavage, F., Makris, F., and Rice, M. (2007). SD/LEP Inclusions/Exclusions in NAEP: An Investigation of Factors Affecting SD/LEP Inclusions/Exclusions in NAEP. Technical report, American Institutes for Research.

Wise, L. L., Hoffman, R. G., and Becker, D. E. (2004).Testing NAEP Full Population Estimates For Sensitivity to Violation of Assumptions. Technical report TR-04-27, Human Resources Research Organization.

Wise, L. L., Le, H., Hoffman, R. G., and Becker, D. E. (2006).Testing NAEP Full Population Estimates For Sensitivity to Violation of Assumptions: Phase II Final Report. Unpublished report, Human Resources Research Organization.

Appendix A: Summary of the Current Modeling Framework

The CMF imputes plausible values for students who are:

1. Selected for NAEP
2. Identified as SD or ELL
3. *Excluded from taking NAEP*.

The data used to fit the model come from students who are:

1. Selected for NAEP
2. Identified as SD or ELL
3. *Included in NAEP, possibly with accommodation*.

The framework is described in McLaughlin (2005). An extended model in Braun, *et al.* (2006) contains one additional, school-level achievement attribute.

Briefly, the framework employs a linear regression in which the response is mean student-level NAEP plausible values, and the predictors are

- Student attributes from ELL or SD questionnaires created by NCES and distributed to schools by NAEP contractors
- School attributes from the CCD or PSS

The fitting procedure employs variable selection. Models are fit separately to data from students who are either ELL or both SD and ELL and from students who are only SD.[11]

Separate models are used to estimate variances, and imputed plausible values are then created by simulation.

The models in McLaughlin (2005) and Braun, *et al.* (2006) also differ, possibly not significantly, with respect to variable selection—specifically, the number of regression models underlying the selection, the treatment of sampling error, and the treatment of missing predictors.

[11] In earlier studies, McLaughlin instead used SD including SD + ELL and ELL alone.