

Linking 2005 NAEP Science Assessments through Bridge Samples¹

--How to design and analyze a trend assessment with context changes

Jiahe Qian

ETS, Princeton, NJ

April 2006

¹ The following paper was prepared for presentation at the 2006 IOMW Meetings and is the copyrighted material of Educational Testing Service. The materials are an unpublished, proprietary work of ETS. Any limited distribution shall not constitute publication. This work may not be reproduced or distributed to third parties without ETS's prior written consent. Submit all requests to <http://www.ets.org/legal/copyright.html>. The work reported herein was supported under the National Assessment of Educational Progress (Grant No. R999G50001, CFDA No. 84.999G) as administered by the Office of Educational Research and Improvement, U.S. Department of Education, and ETS research allocation project 882.04. The opinions expressed herein are solely those of the authors and do not necessarily represent those of Educational Testing Service.

The National Assessment of Educational Progress (NAEP)² science assessments measure what American students know and can do in science subject (Allen, Carlson, & Zelenak, 1999; O’Sullivan, Lauko, Grigg, Qian, & Zhang 2003). One purpose is to provide the reliable measurement of trends in the achievement of science over time. Compared with the previous assessments, the 2005 NAEP science assessment has implemented two major alterations: 1) changing the format of test booklets to that suggested by the National Assessment Governing Board (NAGB), and 2) replacing three released blocks by three new blocks of items. To respond to the changes, the new design of the 2005 NAEP science added national-only bridge sample in addition of the operational sample for each grade assessment. The new design is to assure the stability of trend measures under the new booklet design, and to reduce the possible errors due to the changes.

Studies have found that changes in item order and booklet format can affect user responses. The booklet format changes refer to assessments with identical items but different in item block order, response type, font (design), or overall complexity. Researchers have found that the change in answer order in a multiple-choice test could significantly affect responses between otherwise the same tests (Cizek, 1994; Zwick, 1991). A study suggested that changes between response types would affect how students answer test items (Beaton & Zwick, 1990). Therefore, when assessments involve changes, test creators should prevent creating unequivocal tests and avoid bulky changes (Barron & Koretz, 1996). These effects are called context effects (Robert, 1992), which occur when a test adopts changes such as the differences in item order and booklet format. It implies that the local independence assumption of item response theory (IRT; Lord, 1980) fails to hold in practical test. To measure trends in achievement is a difficult task in measurement. The lesson offered by Beaton and Zwick (1990) says, “If you want to measure change, do not change the measure.”

To avoid the confounding effects of context effects on trend, the new design of the 2005 NAEP science assessment added a bridge sample to bridge the changes and prevent the context effects due to changes. The bridge assessment used booklets of the same format as the previous 2000 assessment. Thus the bridge assessment and previous (2000) assessment could be put on the same scale through concurrent calibration and then used a linear transformation to link the scale to the reporting scale (Allen, Donoghue, & Schoeps, 2001; Braun & Holland, 1982). In addition, because the 2005 operational sample and the bridge sample shared the same population, the scale of current operational assessment could be aligned with that of the bridge assessment

2 The National Assessment of Educational Progress (NAEP) is the nation’s only ongoing representative sample survey of student achievement in core subject areas. In 2005, NAEP conducted a national science assessment of fourth-, eighth-, and twelfth-grade students. State-level results were also collected at the fourth and eighth grades within participating states and jurisdictions. Authorized by Congress and administered by the National Center for Education Statistics (NCES) in the U.S. Department of Education, NAEP regularly reports to the public on the educational progress of students in grades 4, 8, and 12. In 1969, NAEP was authorized by Congress to collect, analyze, and report reliable and valuable information about what American students know and can do in core subject areas. Since that time, in what has come to be referred to as the “long-term trend assessment,” NAEP has assessed public- and nonpublic-school students who are 9, 13, and 17 years old. Since 1990, the more recently developed assessments, referred to as the main NAEP, have also assessed public- and nonpublic-school students in grades 4, 8, and 12.

by matching the means and standard deviations of the distributions through the common population. The new design of the 2005 NAEP science needed to employ two linking processes to ensure the scale of the 2005 NAEP science to be linked to the previous assessment invariantly. Thus the new design did not change the measure of the 2005 NAEP science assessments.

1. NAEP Science Assessments and the recent development

1) NAEP Science Assessments

National concern for students' achievement in science has been the impetus for several recent large-scale efforts to measure science knowledge and skills. In 2005, NAEP administered assessments in science to students at grades 4, 8, and 12. All the assessments were based on frameworks developed through a national consensus process. The results, the average scores and achievement-level performance, were reported for subgroups of students defined by various background and contextual characteristics.

The 2005 NAEP science assessment was administered to combined national and state samples of fourth- and eighth-graders. For the twelfth-graders, the assessment was only administered to national-only samples and did not include state-by-state testing in grade 12. The national sample included students attending both public and nonpublic schools, while the state samples included only students attending public schools.

The 2005 NAEP science assessment was the third administration of an assessment based on The NAEP Science Framework. Students' performance on the assessment is described in terms of average scores on a 0–300 scale for each grade and in terms of the percentages of students attaining three achievement levels: Basic, Proficient, and Advanced. The achievement levels are performance standards adopted by the National Assessment Governing Board (NAGB) as part of its statutory responsibilities and describe what students should know and be able to do.

2) Two major changes in the 2005 NAEP Science Assessments

The 2005 NAEP Science Assessment initiated two major alterations, 1) changing the format of test booklets to that suggested by the National Assessment Governing Board (NAGB), and 2) creating three new blocks of items to replace the released blocks.

The rationale for adopting new format of booklets is to increase the efficiency of the NAEP administration and several of the follow-up tasks if all subject areas are assessed in the same booklet format and administered in common, cross-subject sessions. The rationale for adding a national-only bridge sample is to assure the stability of trend measures under the new booklet design, and to reduce the possible errors due to the changes. The format of booklets used in the bridge sample was the same as in the previous 2000 science assessment. The changes in booklet format and replacement of old blocks necessitated the use of bridge sample as part of the 2005 science assessment.

While all NAEP assessments consist of background questionnaires and cognitive items, the arrangement and timing differ across subject areas. The old science booklets have the following instrument designs:

- 4th grade - Two 20-minute cognitive blocks, 7½-minute general background questionnaire, 7½-minute science background questionnaire, 20-minute hands-on block (half of the students);
- 8th and 12th grade - Two 30-minute cognitive blocks, 7½-minute general background questionnaire, 7½-minute science background questionnaire, 30-minute hands-on block (half of the students).

In place of this diversity, the new design of the 2005 NAEP science booklets are standardized for all three grades:

- Two 25-minute cognitive blocks, followed by a 7½-minute general background questionnaire, and a 7½-minute focused background questionnaire.

To implement the new design, the following changes must be completed and their impact assessed: a) The existing science cognitive items must be reconfigured into 25-minute blocks from their current format. b) The order of presentation of the cognitive and background questionnaires must be modified in science assessments. The new, standard assessment booklets would have meaningful simplifications in administration and reduce the time needed for creating the sampling weights used in the analysis, as well as prospective gains in analysis and reporting timelines.

The 2005 science assessment must serve several basic functions: operational data must be reported, new booklet types must be tested, new items must be tried out, and trend equating must be done. Since trend measures are a vital part of the science assessment, whenever changes to NAEP booklet designs are contemplated, measures must be taken to assure that trend measures can be continued. Since context of a test always exerts context effects upon the parameter estimation (Beaton & Zwick, 1990), the item parameter estimates obtained in any particular booklet configuration are not always appropriate for other conceivable configurations. When booklet format changes, to avoid the conversion in IRT modeling, the new science design added a bridge sample to allow linking of the old operational format to the new format. Because the old and new instrument designs would be administered to randomly equivalent samples, the common population linking approach would set two samples on the same scale through the shared equivalent samples. Moreover, the impact of changing the block design could be measured. Because the format of booklets used in the bridge sample was the same as in the 2000 science assessment, the concurrent calibration approach would put the bridge assessment and the 2000 assessment on the same scale.

2. The NAEP Science Assessment Instruments

The 2005 science assessment booklets at grades 4, 8, and 12 consisted of two 25-minute cognitive blocks. In addition, one-half of the students in each school sample conducted a hands-on task and answered questions related to the task. For this, too, the time allotted was 25 minutes

at each grade. In addition to the science questions that students answered, they also responded to background questions that asked them to give information about themselves and their school experiences. For example, students were asked how much time they spent on homework, how often they used a computer, and what science subjects they were currently taking in school.

The information from the science assessments consisted of the responses of students to the items presented in the assessment. The items were constructed to measure performance on sets of objectives developed by nationally representative panels of learning-area specialists, educators, and concerned citizens. A combination of multiple-choice, short constructed-response (scored dichotomously), extended constructed-response (scored polytomously) and cluster items made up the assessment. The constructed-response items in the assessments were professionally scored. According to the framework, the science assessment was designed to measure three interrelated components of science proficiency: earth, physical, and life. The cognitive domain was divided into conceptual understanding, scientific investigation, and practical reasoning. Each question in the assessment was categorized by its content and cognitive domains.

To ensure that the tasks selected to measure each component covered required contents and a range of difficulty levels, it required relative large items in the pool. To reduce student burden, each assessed student was presented only a fraction of the full pool of items through multiple matrix sampling procedures. In addition to matrix sampling, the Balanced Incomplete Block (BIB) design also balanced the order of presentation of the blocks of questions, except for the hands-on blocks, which always appeared in position three of a booklet.

Table 2.1 shows the scale composition by item types for the 2005 NAEP Science Assessments at each grade, and Table 2.2 summarizes the percentages of item types. There are about 57 percent multiple-choice items and about 40 percent polytomously scored constructed-response items. There are only a few of dichotomously scored constructed-response items: 5 items for grade 4, 4 items for grade 8 and 4 items for grade 12.

The format of booklets in the bridge sample was the same as the previous one, which implies that three released blocks were kept in the bridge assessment. One concern is whether security issue arose in the bridge assessment. In the item analysis, the QC of this issue was implemented by comparing of the score changes of the released blocks and non-released blocks. If security became a problem, the scores of the released blocks would increase much faster between 2005 and 2000 than those of the non-released blocks. But there was no evidence to show the case. For example, for grade 4 in Table 2.3, the score right of the items for the released blocks increased 1.1% between 2005 and 2000, but for the non-released blocks increased more at 1.3%. Clearly no security problems were found for grade 8 and grade 12 either.

3. The Scaling IRT Models

In the analysis of science data, depending on item type and scoring procedure, a separate scale was constructed for each component of science proficiency by using three distinct scaling models. Based on IRT models, each is a “latent variable” model, defined separately for each of

the scale components, which expresses respondents' probabilities to achieve certain scores on the items contributing to a scale as a function of a parameter that is not directly observed.

A three-parameter logistic (3PL) model is used for the multiple-choice items (which are scored correct or incorrect). The fundamental equation of the 3PL model defines the probability that a person whose score on a scale is characterized by the *unobservable* variable will respond correctly to item. A two-parameter logistic (2PL) model is used for the short constructed-response items that are scored correct or incorrect. In addition to the multiple-choice and other two-category items, a number of extended constructed-response items are presented in the science assessments. Each of these items is scored on a multipoint scale with potential scores ranging from 0 to 3, from 0 to 4, or from 0 to 5. The extended constructed-response items, also referred to as polytomous items, are scaled using a generalized partial credit model (GPCM; Muraki, 1992).

The PARSCALE computer program (Muraki & Bock, 1997) was used to estimate the item parameters for the national main assessment. For dichotomous multiple-choice and dichotomized constructed-response items, a three-parameter IRT model was used. In concurrent calibration of IRT model incorporating different logistic forms, the items in new blocks and the released blocks were scaled separately within each of the three subscales.

In scaling, the effects of substitution of experiment kits for one hands-on item were also evaluated. The experiment kits, a set of pin oak acorns used for trend block in grade 4, were substituted because the materials were not available in the quantities needed for the 2005 administration. Neither the pin oaks nor the similar water oaks were available in sufficient quantity to meet the needs of the 2005 assessment. Therefore, filberts (which have similar enough physical characteristics to pin oaks and water oaks) were used as a substitution. However, the effect of substitution of the experiment kits was significant. The item did not fit the GPCM in concurrent calibration and showed evidence of functioning differently across assessments. It was treated as separate items for each assessment year.

During scaling, some polytomous items that did not fit the model received some special treatment (e.g., recoding). If a trend item was recoded in the previous scaling, it would be recoded again for the 2005 assessment. Table 3.1 shows the scale composition by item types after scaling treatment at each grade, and Table 3.2 summarizes the percentages of item types after scaling treatment. The polytomously scored constructed-response items reduced to 25 percent, 30 percent and 29 percent at grade 4, grade 8 and grade 12 separately.

The main purpose of IRT analysis is to provide a common scale on which performance can be compared across groups such as those defined by characteristics including gender and race/ethnicity. However, because of matrix sampling and the BIB-spiraling design of blocks, students do not receive all the items of the scale components. Traditional test scores for individual students, even those based on IRT, would lead to biased estimates of population characteristics, such as subgroup means and percentages of students at or above a certain scale-score level. Consequently, NAEP constructs sets of plausible values designed to represent the distribution of performance in the population. The plausible values are generated from random draws from the predictive scale score distributions by NAEP specific statistical procedure for

computing the statistics of interest, such as mean proficiencies for demographic groups. A plausible value for an individual is not a scale score for that individual, but may be regarded as a representative value from the distribution of potential scale scores for the students in a subpopulation with same characteristics and identical patterns of item response. Statistics describing performance on the NAEP science scale are based on the plausible values. Under the assumptions of the scaling models, these population estimates will be consistent, which implies that the estimates approach the model-based population values as the sample size increases.

4. Linking Science Scales

The NAEP science scales obtained from scaling step used to link to previous assessment scales via common population linking procedures. Since the 2005 NAEP science comprised an operational sample and an added national-only bridge sample, its linking procedure employed two processes.

The first process set the bridge sample and the 2000 assessment sample on the same scale by concurrent calibration and then used a linear transformation to link the scale to the reporting scale. Essentially, the 2000 sample and 2005 bridge sample were calibrated together. Data from the two assessments were scaled together in the same PARSCALE run, specifying the samples for each assessment as coming from different populations. In this linking process, for each scale, the mean and standard deviation of the 2000 data from this joint calibration were matched to the mean and standard deviation of the 2000 data as previously reported. This then linked the 2000 data to the previously established scale. The formula of the transformation for the first process can be found in Appendix.

The second process was based on the common population shared by the bridge and operational samples. The common population between the bridge sample and the operational sample was established on randomly equivalent samples from the same populations. It aligned the scale of current operational samples with that of the bridge samples by matching the means and standard deviations of the distributions. The 2005 NAEP science chose common population equating in stead of common item equating. In common population equating, results for two or more samples from the same population are matched to one another when linking the scales. However, in common item equating, items are assumed to be measuring exactly the same thing for two or more populations, despite any differences in context or administration. As mentioned above, in 2005, NAEP science changed booklet format and used three new blocks of items to replace three released blocks. Consequently the common item equating is not a preferred choice for the case. The formula of the transformation for the second process can be found in Appendix.

Since both grade 4 and grade 8 use combined national and state sample, they share identical linking processes as shown in the flowchart of Figure 4.1. Figure 4.2 shows the flowchart of two linking processes for grade 12, which uses national-only samples in assessments. For the subscale of Physical Science of grade 8, Figure 4.3 shows the Q-Q plot of the distributions of two sets of the transformed scores, reported scores and provisional scores, in the first linking process. For the same subscale, Figure 4.4 shows the Q-Q plot of the distributions of two sets of the transformed scores for bridge sample and operational sample in the second linking process. Figures 4.5 and 4.6 are the Q-Q plots of the linking processes for the

subscale of Earth Science of grade 8. Figures 4.7 and 4.8 are the Q-Q plots of the linking processes for the subscale of Life Science of grade 8. By the plots, the transformations were accurately implemented.

5. Conclusion

This paper has introduced the recent progress in the design and analysis of the NAEP Science Assessments. The 2005 NAEP Science Assessment applied two changes, 1) changing the format of test booklets and 2) replacing the released blocks by three new blocks of items. To adapt these changes, the strategy of the 2005 NAEP science design is to add a national-only bridge sample in addition to the operational sample. The decision was made to avoid the confounding effects of the changes in booklet format and item blocks.

The fundamental reason that causing confounding effects is that the assumption of the local independence assumption in IRT models is not always hold in empirical data. Therefore, student responses will be affected when involving changes in tests, such as release of item blocks, modification in booklet format or alteration administration procedures. Under changes in tests, the parameter estimates of IRT models become not robust. The item parameter estimates obtained in any particular booklet configuration are not always appropriate for other conceivable configurations.

Consequently, the procedures of the data analysis should be consistent with new design. Different methods should be used in analyzing data. In the item scaling step of the 2005 NAEP science, three different IRT models were scaled by concurrent calibration. However, the items in the new blocks and the released blocks, including those functioning differently across assessments, were scaled separately within each of the three subscales. In the linking step, the 2005 NAEP science assessment employed two linking processes. The first process set the bridge sample and previous assessment (2000) sample on the same scale by concurrent calibration and then used a linear transformation to link the scale to the reporting scale. The second process used common population linking to align the scale of current operational samples with that of the bridge samples by matching the means and standard deviations of the distributions.

The results of the 2005 NAEP science show that the new design successfully has assured the stability of trend measures under change of the booklets format and reduced the extra measurement errors due to the changes. The new NAEP science design has provided an example to handle changes in large scale assessments and to avoid the consequences of context effects. The idea can be adapted without difficulty to similar situations. The design and analysis of the 2005 NAEP science provided a revision to the lesson offered by Beaton and Zwick (1990): To measure a trend with context changes, bridge the changes but do not change the measure.

Appendix

The linear transformation is used to link the provisional scale, obtained from calibration, to the reporting metric. The transformation matches overall mean and standard deviation of the plausible values to those of the previous reporting metric.

In the first process, let R_B be the set of bridge sample and R_{00} be the set of reporting sample in 2000. Let $\theta_{i \in R_B, provisional}$ be the scores on the provisional scale for case i in the bridge sample and $y_{i \in R_B, reporting}$ be the transformed scores on the reporting scale for case i in the bridge sample. The formula of the transformation is

$$y_{i \in R_B, reporting} = A_1 * \theta_{i \in R_B, provisional} + B_1,$$

where $A_1 = S_{R_{00}, 00, reporting} / S_{R_{00}, 05, provisional}$ and $B_1 = \bar{y}_{R_{00}, 00, reporting} - A_1 * \bar{\theta}_{R_{00}, 05, provisional}$. The symbol $S_{R_{00}, 00, reporting}$ is the standard deviation of the 2000 reporting scores for the cases in R_{00} and $S_{R_{00}, 05, provisional}$ is the standard deviation of the 2005 provisional scores for the cases in R_{00} ; $\bar{y}_{R_{00}, 00, reporting}$ is the mean of the 2000 reporting scores for the cases in R_{00} and $\bar{\theta}_{R_{00}, 05, provisional}$ is the mean of the 2005 provisional scores for the cases in R_{00} .

In the second process, let R_B be the set of bridge sample and $R_{EQ, B}$ be the randomly equivalent sample of the common population in R_B . Let R_{05} be the set of the 2005 reporting sample and $R_{EQ, 05}$ be the equivalent common population in R_{05} . So $R_{EQ, B} \subset R_B$ and $R_{EQ, 05} \subset R_{05}$. Let $\theta_{i \in R_{05}, provisional}$ be the scores on the provisional scale for case i in R_{05} and $y_{i \in R_{05}, reporting}$ be the transformed scores on the reporting scale for case i in R_{05} . The formula of the transformation is

$$y_{i \in R_{05}, reporting} = A_2 * \theta_{i \in R_{05}, provisional} + B_2,$$

where $A_2 = S_{R_{EQ, B}, 05, reporting} / S_{R_{EQ, 05}, 05, provisional}$ and $B_2 = \bar{y}_{R_{EQ, B}, 05, reporting} - A_2 * \bar{\theta}_{R_{EQ, 05}, 05, provisional}$. The symbol $S_{R_{EQ, B}, 05, reporting}$ is the standard deviation of the 2005 reporting scores for the cases in $R_{EQ, B}$ and $S_{R_{EQ, 05}, 05, provisional}$ is the standard deviation of the 2005 provisional scores for the cases in $R_{EQ, 05}$; $\bar{y}_{R_{EQ, B}, 05, reporting}$ is the mean of the 2005 reporting scores for the cases in $R_{EQ, B}$ and $\bar{\theta}_{R_{EQ, 05}, 05, provisional}$ is the mean of the 2005 provisional scores for the cases in $R_{EQ, 05}$.

References

- Allen, N., Carlson, J., & Zelenak, C. (1999). The NAEP 1996 technical report (NCES 1999-452). Washington DC: National Center for Education Statistics.
- Allen, N., Donoghue, J., & Schoeps, T. (2001). The NAEP 1998 technical report (NCES 2001-509). Washington DC: National Center for Education Statistics.
- Beaton, A. E., & Zwick, R. (1990). *The effect of changes in the National Assessment: Disentangling the NAEP 1985-86 reading anomaly*. (No. 17-TR-21) Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.
- Barron, S. I. and Koretz, D. M. (1996). An Evaluation of the Robustness of the National Assessment of Educational Progress Trend Estimates for Racial-Ethnic Subgroups. *Educational Assessment*, 3(3), 209-248.
- Braun, H. I., & Holland, P. W. (1982). Observed score test equating: A mathematical analysis of some ETS equating procedures. In *Test Equating*. (P. W. Holland and D.B. Rubin, eds.) Academic Press: Princeton, New Jersey.
- Cizek, G. J. (1994). The Effect of Altering the Position of Options in a Multiple-choice Examination. *Educational and Psychological Measurement*, 54(1): 8-20.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2), 159-176.
- Muraki, E., & Bock, R. D. (1997). *PARSCALE: IRT item analysis and test scoring for rating-scale data*. Chicago, IL: Scientific Software International.
- O'Sullivan, C.Y., Lauko, M.A., Grigg, W.S., Qian, J., and Zhang J. (2003). *The Nation's Report Card: Science 2000* (NCES 2003-453). Washington DC: National Center for Education Statistics.
- Robert L. B. (1992). The Context of Context Effects. *Applied Measurement in Education*, Vol. 5, No. 3, 225-264.
- Zwick, R. (1991). Effects of Item Order and Context on Estimation of NAEP Reading Proficiency. *Educational Measurement: Issues and Practice*, 10 (3), 10-16.

Table 2.1 Scale composition by item types
for the 2005 NAEP Science Operational Samples
(With no adjustments)

Subscale	Multiple-choice items	Dichotomously scored constructed-response items	Polytomously scored constructed-response items
Grade 4			
Physical	19	4	22
Earth	35	0	19
Life	33	1	19
Total	87	5	60
Grade 8			
Physical	38	2	27
Earth	42	1	27
Life	46	1	39
Total	126	4	93
Grade 12			
Physical	47	1	28
Earth	38	3	29
Life	30	0	27
Total	115	4	84

Table 2.2 Percentages of Item Types
for the 2005 NAEP Science Operational Samples
(With no adjustments)

	Multiple-choice items (MC)	Dichotomously scored constructed-response items (2PL)	Polytomously scored constructed-response items (GPCM)
Grade 4	57.2	3.3	39.5
Grade 8	56.5	1.8	41.7
Grade 12	56.7	2.0	41.4

Table 2.3 Comparison of the Percentages Right of Released Blocks and Non-released Blocks

Grade	Change in % Right 2005-2000		
	Released Blocks (1)	Non-released Blocks (2)	(2) - (1)
4	1.13	1.31	0.18
8	-1.24	-0.04	1.20
12	-0.03	1.03	1.06

Table 3.1 Scale composition by item types
for the 2005 NAEP Science Operational Samples
(with adjustments in scaling)

	Multiple-choice items (MC)	Dichotomously scored constructed-response items (2 PL)	Polytomously scored constructed-response items (GPCM)
Grade 4			
Physical	19	10	16
Earth	35	9	10
Life	33	8	12
Total	87	27	38
Grade 8			
Physical	38	10	19
Earth	42	9	18
Life	46	9	30
Total	126	30	67
Grade 12			
Physical	47	9	20
Earth	38	13	19
Life	30	7	20
Total	115	29	59

Table 3.2 Percentages of Item Types
for the 2005 NAEP Science Operational Samples
(with adjustments in scaling)

	Multiple-choice items (MC)	Dichotomously scored constructed-response items (2PL)	Polytomously scored constructed-response items (GPCM)
Grade 4	57.2	17.8	25.0
Grade 8	56.5	13.5	30.0
Grade 12	56.7	14.3	29.1

Figure 4.1. 2005 Science: Grades 4 & 8 Linking Design

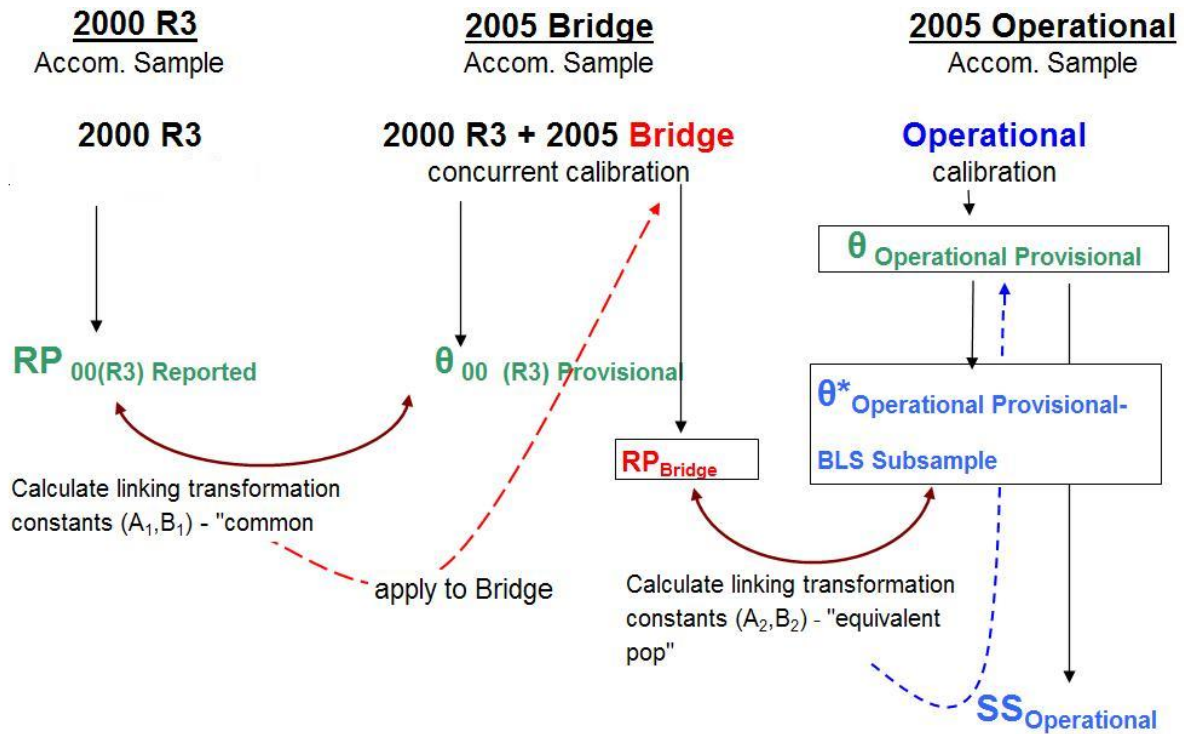
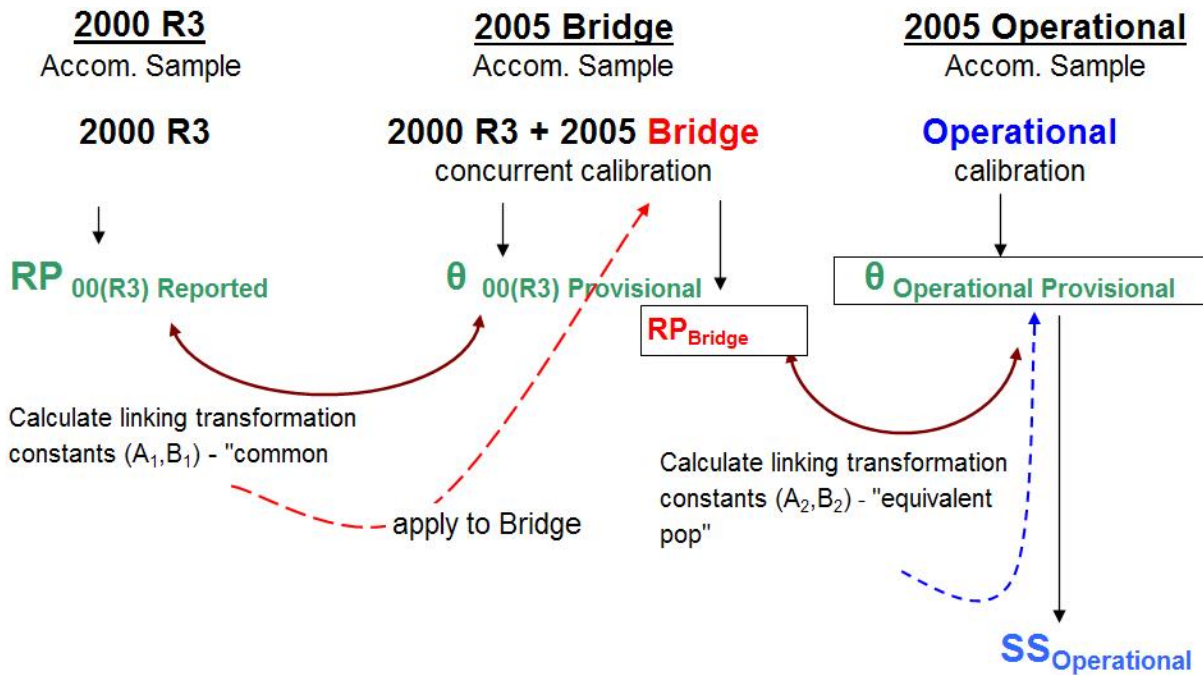
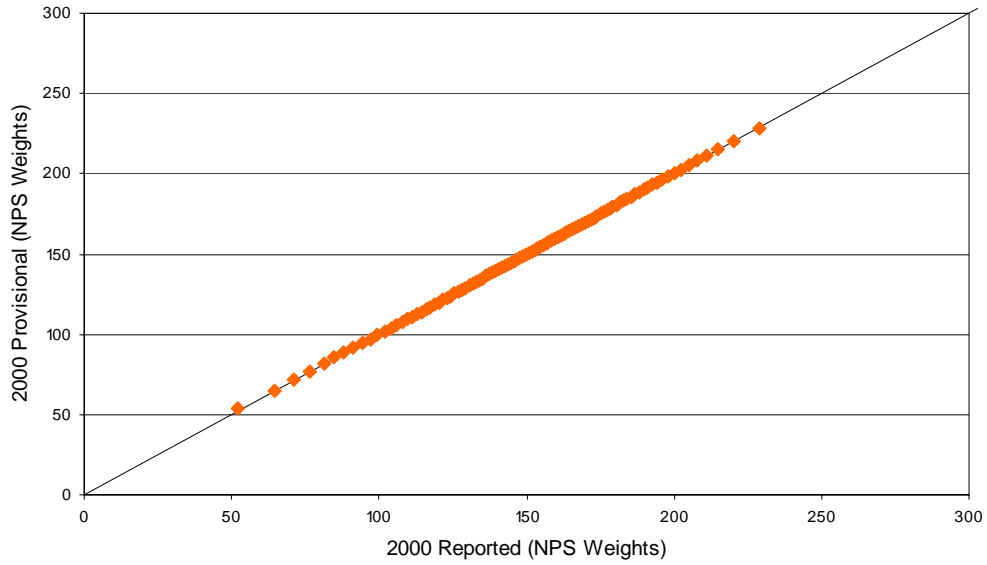


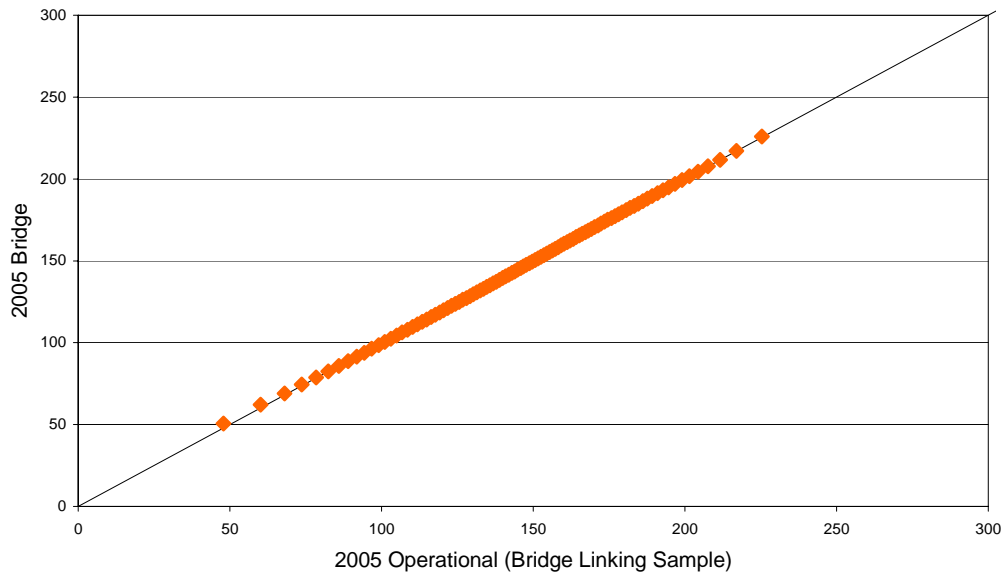
Figure 4.2. 2005 Science: Grade 12 Linking Design



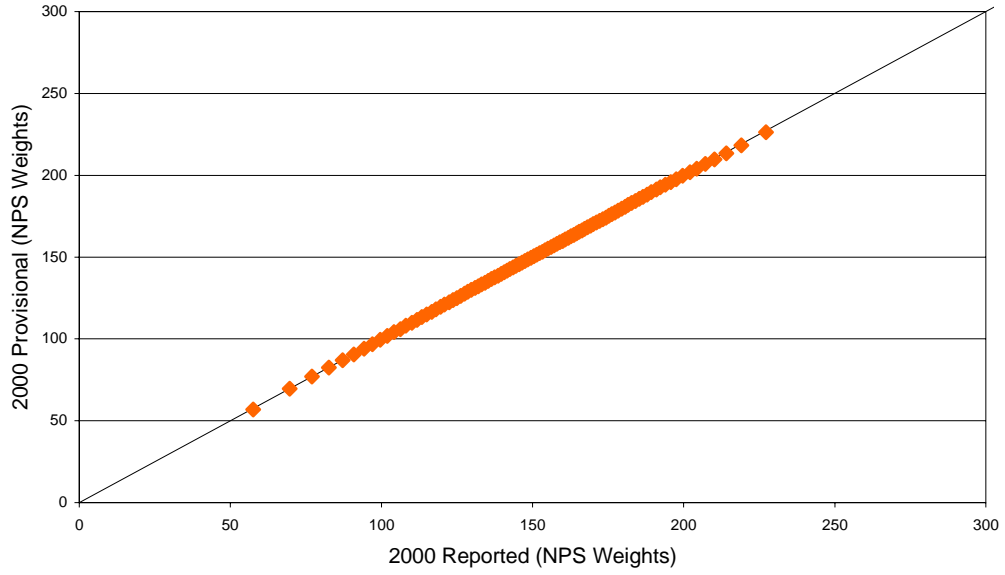
**Figure 4.3. 2005 NAEP Science Grade 8:
Linking of 2000 Distributions
Scale 1 - Physical Science**



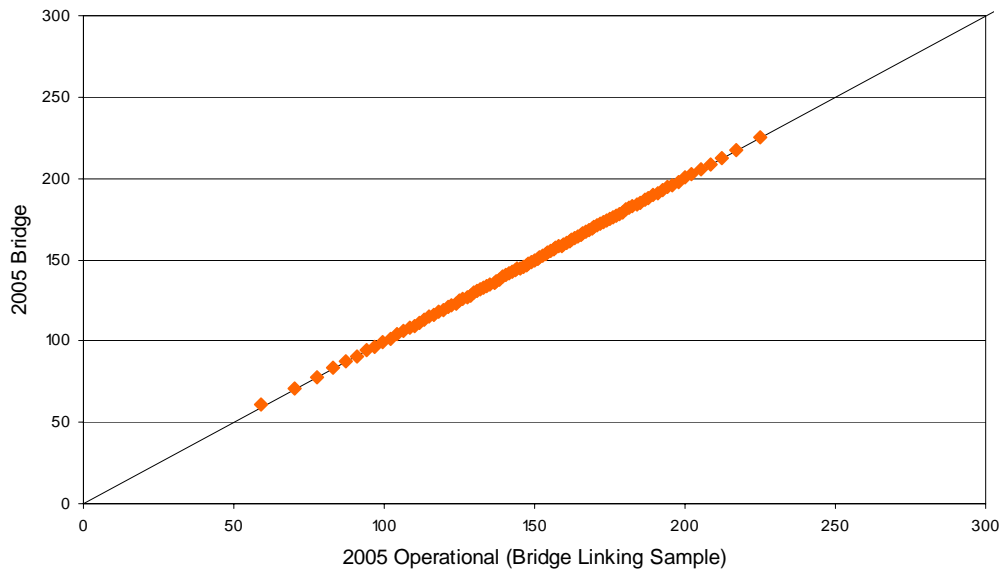
**Figure 4.4. 2005 NAEP Science Grade 8:
Linking of 2005 Distributions
Scale 1 - Physical Science**



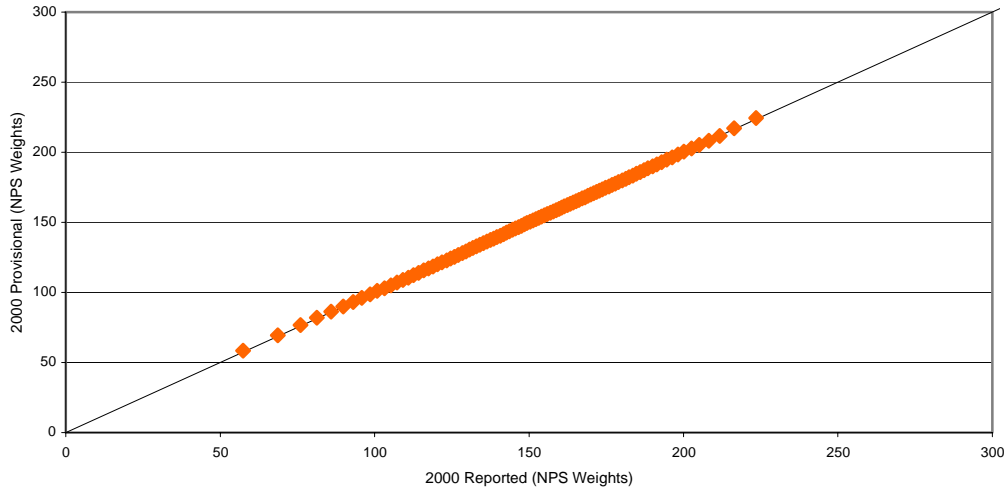
**Figure 4.5. 2005 NAEP Science Grade 8:
Linking of 2000 Distributions
Scale 2 - Earth Science**



**Figure 4.6. 2005 NAEP Science Grade 8:
Linking of 2005 Distributions
Scale 2 - Earth Science**



**Figure 4.7. 2005 NAEP Science Grade 8:
Linking of 2000 Distributions
Scale 3 - Life Science**



**Figure 4.8. 2005 NAEP Science Grade 8:
Linking of 2005 Distributions
Scale 3 - Life Science**

