A Retrospective Linking of ECLS-K and ECLS-B Reading Scores

Carolyn G. Fidelman

Early Childhood and Household Studies
National Center for Education Statistics
carolyn.fidelman@ed.gov

This discussion is intended to promote the exchange of ideas among researchers and policy makers. The views expressed here are part of ongoing research and analysis and do not necessarily reflect the position of the U.S. Department of Education.

Abstract

Plans for making scores between different cohorts of examinees verifiably construct-equivalent are ideally built into a test's design from the beginning. But it is sometimes in hindsight that we realize we might be able to leverage the results from two very similar studies in order to carry out new comparisons. In cases where test forms for different cohorts are very similar yet have not been thoroughly tested for equivalence before going to the field, can we reformulate the scores in order to achieve a degree of parity adequate for score linking? The psychometric properties of the items of two multistage kindergarten reading tests written to similar specifications and administered to cohorts eight years apart, in 1998 and 2006, were examined and a subset of the test items was used in order to meet the assumptions of item response theory parameter estimation and score linking. Specifically, a unidimensional set of common items was identified from both stages of the tests and the unique items measuring that same construct within and across the two cohorts were able to be put on the same scale. IRT concurrent calibration within and chain linking across cohorts was found to be the best approach for linking the scores of Early Childhood Longitudinal Study (ECLS) 1998 and 2006 Reading assessments. By putting the scores of the two tests on one scale, we can offer new versions of the participant scores for use in studies designed to compare or contrast the two cohorts. The results from this linking study also remind us of some common ways in which construct validity can be verified and maintained over multiple cohorts or how modifications in structure can be made while maintaining the baseline measurement properties that enable their continued use in such analyses.

Introduction

This paper describes an analysis being conducted at the National Center for Education Statistics (NCES), which is part of the Institute of Education Sciences within the U.S. Department of Education. Its purpose is to achieve comparability of test scores for academic achievement instruments used in the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K) and Early Childhood Longitudinal Study, Birth Cohort of 2001 (ECLS-B). This paper will deal with one content area of that analysis which is linking the scores for *reading* assessments of first time kindergartners in 1998 and 2006. The larger analysis investigates linking of both reading and math scores and provides much more detail than will be found in this paper. The word ―retrospective" in the title refers to the fact that linking assessment results of the two studies was not originally built into the assessment design of either, but that we can work with the item level data and we can retrofit them for a new version of the scores that are comparable across the studies. An important goal of the study is to show how we can leverage data that we already have and make it useable for new types of analyses.

About the ECLS

The Early Childhood Longitudinal Studies, or ECLS, are sponsored by NCES (see the references section of this paper for a listing of relevant publications.) The purpose of the first ECLS study, ECLS-K, is to provide detailed information about the academic, social, and physical development of students from kindergarten entry to middle school. Westat, Inc. and NORC assisted with the design of the study and Westat, Inc. collected the data upon which this study is based. The students participating in the ECLS-K were followed longitudinally from kindergarten in the fall of 1998 through the spring of 2007 when most were in eighth grade. The reading scores discussed in this paper

are based on data from a subset of about 3,000 students who were first-time kindergartners in the fall of 1998, who took both the reading and mathematics assessments, and for whom demographics data are present. The original sample included approximately 21,000 kindergartners. Subsetting the sample was done to facilitate score development for cross-study purposes and to balance the ECLS-K sample by ability level. There were 72 unique items in the original reading assessment instrument, administered in varying combinations across three ability levels. Item parameters derived from these results are applied to all 21,000 ECLS-K sample members in final rescoring. When properly weighted, information about these students represents all students who were in kindergarten in the 1998-99 academic year. More information about the ECLS-K data and publications can be found at the ECLS-K website, http://nces.ed.gov/ecls/kindergarten.asp.

A second ECLS, the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), also sponsored by NCES, studied children beginning at age 9 months and through kindergarten entry and is representative of the 3.9 million children born in the United States in 2001. RTI and Westat, Inc. contributed to the design and conduct of ECLS-B. The focus of the ECLS-B is children's early development; their home learning experiences; their experiences in early care and education programs; their health care, nutrition, and physical well-being; and how their early experiences relate to their later development, learning, and success in school. As with the ECLS-K data, the approximately 3,000 cases selected for the current linking analysis were drawn from examinees who took both the math and reading assessments, for whom demographic data were present, and were balanced by ability level. Approximately 7,000 children participated in ECLS-B during the 2006 collection, when most of the sample was in kindergarten. As with subsetting employed with the ECLS-K data, the ECLS-B sample was subset to facilitate comparable scores between the two studies. Balancing across achievement levels was minimal for the ECLS-B data because of the way the study was designed. In consideration of the fact that ECLS-B children were all between the ages of 57 and 75 months at age of first kindergarten testing, ECLS-K cases used in the present study were selected to match that age range as well. There were 62 unique items in the original reading assessment instrument, administered in varying combinations across three ability levels. More information about the ECLS-B data and publications can also be found at the ECLS-B website, http://nces.ed.gov/ecls/birth.asp.

Goals of the Study

The data from the ECLS studies has been used by a variety of researchers (see http://nces.ed.gov/ecls/pdf/bibliography.pdf for the kinds of studies that have been conducted). While most of these studies use data from only one or the other study, researchers have sought approaches that would allow use of the data from both studies in a comparative way.

In light of the many developments in measurement theory and in the technical aspects of test scoring since the 1990s when the first study was designed, it is worth revisiting the basic scoring approach. Because of the high cost of these studies, we need to make sure that we are getting the best information possible from them. While it is convenient or possibly necessary to use somewhat imperfect designs for feasibility reasons, we want to make those decisions in a conscious way, either in our use of the extant data or in planning for new data collections. Not only do we want the most precise scores possible, we may discover (as we did) that cost savings could result where we find that we need fewer cases than specified in the original designs at least in terms of test development. (The notion of the sample needed for test development should be kept separate from that of the main study sample requirements, which are driven by entirely different requisites.) Thus we are obliged to review the scoring methodology of the original studies to confirm that they work or to make adjustments. The existence of the original raw data permits us to revisit these decisions and possibly propose some adjustments or a prototyping of new versions of the scores.

The success of ECLS-K and ECLS-B demonstrated the feasibility of launching a new kindergarten cohort study that began in 2011. The new study is intentionally similar in design and scope to the original ECLS-K. As we begin to envision this and a succession of such studies it is natural that we will want to consider not only the changes within the cohorts, but also the comparisons across cohorts. We will need to establish the broader vision of education and academic development under which the reading and math tests are being designed in order to create and maintain comparability.

With the above in mind, we propose four goals for this study.

1. Explore how a second ECLS study that includes kindergarten-year assessments can be meaningfully compared to the first kindergarten cohort study.
2. Demonstrate the feasibility of producing scores for subsequent ECLS studies that are on the same scale as the 1998 ECLS-K study.
3. Produce revised score sets for ECLS-K 1998 Round 1 and ECLS-B Round 4 for use in comparison studies.
4. Suggest an approach to score linking that can be incorporated into future data collection designs.

Data Preparation

Before parameter estimation and before linking the scores of the ECLS-K kindergarten reading test to that of the ECLS-B, the author had to investigate and understand the data structures of ECLS-K and ECLS-B reading assessments and how they related the original scoring methodologies.

Due to the use of concurrent calibration to link scores in the original studies, there was no need in the original scaling for identification of examinees by level in the item level data set. However, the author wanted to qualitatively assess the similarity of the level tests and also leave open the possibility of employing a linking methodology that requires separation of the data into the three ability level tests. It was therefore necessary to determine which items were common across all tests and which items were unique to the various level tests. Another change had to do with item naming. Identical items across the two studies had different names. For the present analysis a new common name was assigned to the identical items after investigating their equivalence in terms of wording, formatting, ordering and administration protocol. This enabled the author to construct a dataset that merged all items and examinees over both cohorts. With the use of child ID, cohort, and level variables, this dataset could be used to evaluate a variety of analysis approaches. Table 1 shows a partial list of the items and their instance across ECLS-K and ECLS-B.

Table 1. Item Content and Sequence Equivalency Chart for ECLS-K & ECLS-B Kindergarten Reading

					Order of Item Presentation					
Revised item name	Dim98	K-R1 Code	B - R4 Code	Content	K Low	K Mid	K High	B Low	B Mid	B High
KBR01	LR	XYZLN	LL39	[content removed]	*1*	*1*	*1*	*1*	*1*	*1*
KBR02	LR	XYZME	LL45	[content removed]	*2*	*2*	*2*	*3*	*3*	*3*
KBR03	LR	XYZPJO	LL48	[content removed]	*3*	*3*	*3*	*4*	*4*	*4*
KBR04	LR	XYZWP	LL42	[content removed]	*4*	*4*	*4*	*2*	*2*	*2*
KBR05	BS	JKLS	LL51	[content removed]	*5*	*5*	*5*	*5*	*5*	*5*
KBR06	BS	JKLC	LL54	[content removed]	*6*	*6*	*6*	*6*	*6*	*6*
KBR07	BS	JKLQ	LL57	[content removed]	*7*	*7*	*7*	*7*	*7*	*7*
KBR08	BS	JKLG	LL58	[content removed]	*8*	*8*	*8*	*8*	*8*	*8*
KBR09	ES	JKPS	LL72	[content removed]	*9*	9	9		*12*	12
KBR10	ES	JKPE	LL69	[content removed]	*10*	10	10		*11*	11
KBR11	ES	JKPZ	LL60	[content removed]	*11*	11	11		*9*	9
KBR12	ES	JKPD	LL63	[content removed]	*12*	12	12		*10*	10
KBR13	SW	WERR	LL75	[content removed]	*13*	13	13		*13*	13
KBR14	SW	DECV	LL81	[content removed]	*14*	14	*14*		*15*	15
KBR15	SW	XSRT	LL78	[content removed]	*15*	15	15		*14*	14
KBR16	SW	HGND	LL84	[content removed]	*16*	16	*16*		*16*	16
KBR17	IU	SDNFD	LL87	[content removed]		17	*17*			*17*
KBR18	IU	DHEGF	LL93	[content removed]		18	*18*			*19*
KBR19	IU	SJEFGD	LL90	[content removed]		19	*19*			*18*
KBR20	IU	WAPLK	LL96	[content removed]		20	*20*			*20*
KBR21	PF	DEPOL		[content removed]	*17*					
KBR22	PF	LOPIE	LL108; LL141	[content removed]	*18*			*12*	*22*	
KBR23	PF	SOPIR	LL99; LL126	[content removed]	*19*	*21*	21	*9*	*17*	
KBR24	PF	LOERN	LL102; LL129	[content removed]	*20*	*22*	22	*10*	*18*	
KBR25	PF	SOPIN	LL105; LL132	[content removed]	*21*	*23*	23	*11*	*19*	
KBR26	PV	CKLWP		[content removed]	*22*					

NOTE: Items in italics in the last six columns were found via factor analysis to belong to the main general reading construct being tested
SOURCE: Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort, Kindergarten-Year Reading Assessment. National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Cohort, Kindergarten-Year Reading Assessment.

Table 1 merits extended explanation. This table shows a sample of 25 of the total of 92 items identified as being used in either or both of kindergarten-year reading assessments in ECLS-K and ECLS-B. Of the 63 ECLS-K items and 72 ECLS-B items, 42 were common to both assessments for one or several level tests. Items 1-20 listed in the table were used in the original stage 1 or router test. This router test helped administrators to decide which of three stage 2 tests to administer the child. The first column contains the revised item names. Parameters for italicized items in column 1 were as estimated for ECLS-K and applied to ECLS-B in the linking. Items in plain text in column 1 remained unanchored for the calibration because they were present in one cohort only. Note, of the 92 items considered, 18 were removed from the reading assessment item pool because they did not load efficiently on the primary reading dimension identified during factor analyses conducted for this analysis . A blank cell in one the last six columns indicates that that item was not part of that level test. The numbers in the last six columns indicate the order that an item was presented in a level test. The second column, *Dim98*, codes the item for the construct that it represented according to the ECLS-K test specifications. *LR* is letter recognition; *ES* is ending sound; *SW* is sight word; *IU* is initial understanding; *PF* is print familiarity.

Assessing Model Fit and Subsetting the Sample

The original methodology involved applying a 3 parameter IRT model to the data. This model was applied to each round of data for a domain test such as reading and added the data of any previous rounds to that year's round in order to put all scores on the same scale. Thus a new set of scores for the current year and previous years or rounds was published for each year of the study. This accretive approach using concurrent calibration makes it such that the final version of the data published with the final year of the study is considered the best, most accurate data set for researchers performing analyses on ECLS-K alone when including the academic test scores in their analyses (see http://nces.ed.gov/ecls/dataproducts.asp). More information about the original scoring methodology is documented in *Early Childhood Longitudinal Study, Kindergarten Class of 1998 - 99 (ECLS-K), Psychometric Report for the Eighth Grade* (NCES 2009-002). Please note that any new scores produced for this analysis are constructed for the main goal of enabling comparability studies. This does not imply that the currently published scores are any less valid for analyses involving only the ECLS-K or ECLS-B study data.

In constructing data to facilitate comparisons across ECLS-K and ECLS-B, the author revisited the scoring methodology, recalibrating and rescoring the item level ECLS-K 1998 data by applying some basic principles of item response theory as proposed by Hambleton, et al. (1991). There are four basic assumptions of IRT, that is, things that we need to ascertain about the data before deciding on a particular IRT model or whether IRT methodology is appropriate at all.

1. Does the test measure one thing?
2. Is there equal discrimination across items?
3. Is guessing minimal or absent?
4. Is the test non-speeded?

In producing both the reference (1998) and focal (2006) scores I checked the data for unidimensionality, that is, that the test measures one thing. These tests were designed to measure a general reading ability, much in the way that NAEP does. I used factor analysis by each of the three ability level tests to detect which items contributed to the main first factor represented by the test at that level. Some items loaded well on the first factor at one ability level but not at another. This was not necessarily a reason to drop an item. But the 18 items that did not load efficiently on the main factor at any ability level were dropped from the analysis. An item's unidimensionality rating is reflected in the italicization of items in the last six columns of table 1 as explained in the note below the table.

A 1 parameter IRT model assumes that the items all discriminate equally between examinees that have the target ability and those that do not. Where item discrimination varies considerably, a minimum 2 parameter model is indicated. If empirical analyses indicate that guessing is a factor, then one would apply the 3 parameter IRT model. To determine which model to use we consider the testing context qualitatively and also perform a series of data checks.

The item discrimination or point biserial values for these items varied quite a bit, from .04 to .73, a strong argument for a 2 parameter model. Regarding guessing, in this test there was no time limit and children were given ample time to answer. Also, administrators were allowed to stop the test if there was a certain level of frustration. Given those aspects of the testing protocol it was possible that a 2 parameter model would be adequate. On the other hand, test administrators reported that children appeared to be guessing at times. One could also argue that, inasmuch as a young child wants to please an adult test administrator, the situation might be considered somewhat high stakes. High stakes testing contexts are often assumed to best fit the 3 parameter IRT model. This led me to need to do other examinations of these data to determine whether a 2 or a 3 parameter model would be a better fit.

Model fit in IRT is not unlike the appropriate selection of a model for performing regression analysis. One would not apply ordinary least squares linear regression on data that was curvilinear or log linear in nature. If we had the entire population of examinees and could give them the test many times, erasing their memories of taking the test each time, we could detect a clear ogive fit line that reflected the true nature of the probabilities of correct responding by children at all the various ability levels. Instead we usually have noisy data from imperfectly

specified and executed testing situations. In item response theory methodology, we evaluate fit to a 1P, 2P or 3P model according to the characteristics of the data we have. The best model's fit lines or item characteristic curves will show the fewest unexpected results from the real test data.

The example shown in table 2 is the result of a fit analysis of the ECLS-K reading test. Using the unidimensionality criterion, we selected 62 items and concurrently calibrated them using the IRT software PARSCALE 4.1 (Muraki & Bock, SSI, Inc. 2003). The parameter estimation was performed under six different conditions: two sample types and three IRT models. For the sample variations, either all examinees were used or a randomly selected subset of examinees was chosen, balanced by test levels in the reading assessments. This is because we notice that, in the raw data for these tests, quite different numbers of examinees took the various level tests. The data became much better behaved once a rule suggested in duToit (2003) was applied that reduced the number to approximately 1,000 per level test. As du Toit points out there is not much precision to be gained for any n greater than 1,000. And the fact that the ECLS-K low ability level test had over 12,000 examinees compared to 3,000 for the middle ability level and 600 for the high ability level suggested that balancing the sample for purposes of comparison to ECLS-B would be appropriate. All models performed better when the balanced, randomly subset samples were used.

As for a choice of IRT models, at least two tests on the data showed that a 2 parameter IRT model was the best fit for this data. First, one can see in table 2 that the -2 Log likelihood differences that are obtained at the end of the expectation-maximization (EM) algorithm[1] that is applied in the estimation process in PARSCALE between models obtained significantly lower values for the 2P model than for either the 1P or 3P models. A lower -2LL says that this model, with its estimation of what the true parameters might possibly be, is more likely than another with a higher -2LL. This -2LL is chi-square distributed on a number of degrees of freedom equal to the number of items in the test. The significance of the difference between two models can be found using a chi-square table. The critical value for the difference between two models here is 81 (for df=62). The difference between the 3P and 2P models for the subset sample is 26,791 (129,378 – 102,587), which is well over the chi-square critical value for this data. This indicates superior fit of the 2P model.

As another check, one sees in the last column of table 2 that many more items satisfied the chi-square test for model fit under the 2P model than under the other models. There are other methods that could be applied in assessing IRT model fit, such as analysis of residuals. Metrics shown here are intended to illustrate why a 2P model was selected for this analysis. Other metrics generated similar results. The fit tests reported here were also performed on the ECLS-B data and similar results were obtained. The final decision was to use a 2 parameter IRT model and to use the subset data, randomly selected and approximately balanced by level test n, while satisfying the need for examinees to have scores on both a math and a reading test, to be between the ages of 57 and 75 months, and to have demographic data that included age, sex, ethnicity, SES, region, mother's age, number of siblings, urbanicity, and home language. These last variables were used to verify the comparability of the development samples on what we considered some very basic characteristics and will be shown in a subsequent report currently under development.

[1] An expectation–maximization (EM) algorithm is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables.

Table 2. ECLS-K Reading Model Fit (for 62 items)

	Number of cases	Cases per level Low	Mid	High	Difficulty	*sd*	Discrim	*sd*	converge	-2LL	% items fit
1P	16450	12550	3300	600	1.12	*1.67*	1.00	*0.00*	yes	591526	8%
1P	3000	1250	1150	600	0.20	*1.37*	1.00	*0.00*	yes	106263	26%
2P	16450	12550	3300	600	1.20	*1.79*	0.91	*0.55*	yes	591906	0%
2P	3000	1250	1150	600	0.03	*1.48*	1.38	*0.79*	yes	102587	69%
3P	16450	12550	3300	600	0.38	*1.06*	2.17	*1.64*	no	594160	0%
3P	3000	1250	1150	600	0.38	*1.12*	1.72	*0.92*	no	129378	47%

NOTE: *sd*=standard deviation; *Difficulty* is the average IRT difficult parameter for the test and generally ranges from -4 to +4; *Discrim* is the average IRT discrimination parameter for the test and generally ranges from 0-4; *-2LL* is the -2 log likelihood of the model at the last EM cycle.
SOURCE: Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort, Kindergarten-Year Reading Assessment public use data set (NCES 2010-010). National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Cohort, Kindergarten-Year Reading Assessment public use data set (NCES 2009-005).

Score Linking

When linking the scores of two tests there are several things one should check in order to meet comparability requirements set forth in Kolen & Brennan (2004, pp. 269-275).

1. Are the ECLS-K and ECLS-B tests adequately comparable forms from the framework and design standpoint for the purposes of linking?

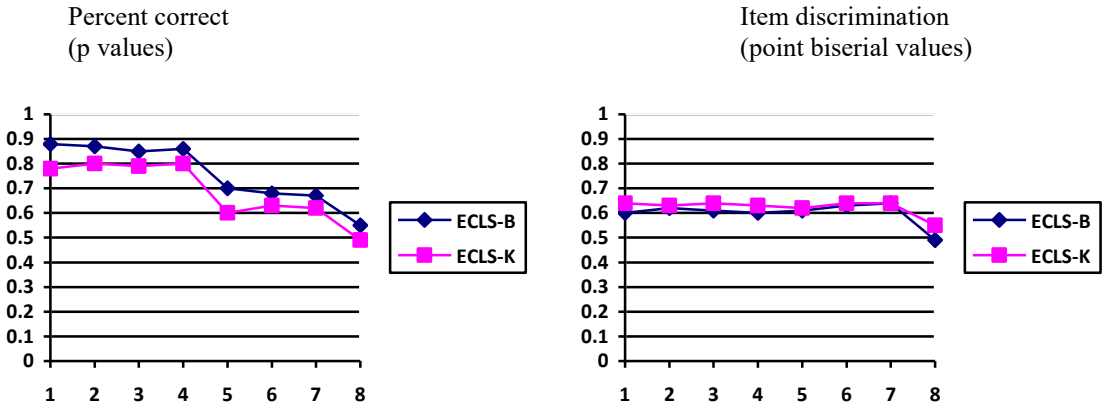
2. Does the unidimensionality property hold across years of a given domain test?

3. To what extent do the common item sets of the tests have the following qualities?
    a. They are 20% or more of total test items.
    b. Their content represents the same content and in the same proportions as the larger test.
    c. They are identically worded and formatted across tests.
    d. They are identically ordered or placed across tests.
    e. They were administered using the same protocol.
    f. They match statistically.

4. To what extent do the entire item sets of the tests have the following qualities?
    a. Their item p-values are equally distributed.
    b. The tests are equally reliable.

A qualitative analyses of the design, specifications, and item distribution by content showed the tests to satisfy the framework similarity requirements. While unidimensionality within the tests was verified with factor analysis, it was not possible to do this across cohorts within the time constraints of this study. At this time, we rely on the stated intent and design of the reading tests of the two cohorts to reassure us that the two tests measure the same reading construct broadly defined. Figure 1 shows results of a check of requirements 4.a. and 4.b. from the list above.

Figure 1. Comparison of Item Difficulty and Item Discrimination for the 8 Common Items of the ECLS-K and ECLS-B Kindergarten Reading Tests



SOURCE: Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort, Kindergarten-Year Reading Assessment public use data set (NCES 2010-010). National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Cohort, Kindergarten-Year Reading Assessment public use data set (NCES 2009-005).

On the left side of figure 1 is a plot of the percent correct or p values for the two tests. They map very closely as do the point biserial values shown on the right side of figure 1. Overall, it appears that the data from these two tests are adequately similar for linking.

The Linking Type

It is important to understand that certain claims are associated with the various terms in common usage for the activity that is generally called *linking*. Below are the various labels given to the types of linking, in order from most to least constrained:

1. Equating
2. Calibration
3. Statistical moderation
4. Projection
5. Concordance[2]

With *equating* we are talking about calibrating interchangeable forms of a test. They are designed exactly to the same test blueprint. That means they have the same number of items and the content of the items is matched. For example, the Defense Language Aptitude Battery comes in Forms A, B, C, & D. It should be indifferent to the examinee as to which form she takes. *Calibration* is where the items are written to the same test blueprint but the length of the test differs or selection of items targets a different ability level. *Statistical moderation* involves linking test forms that are designed to the same construct but not the same exact test framework. It follows that the test specifications of the two instruments are invariably different in the case of a statistical moderation. The number of items may differ and the distribution of the content over items may differ somewhat. *Projection* is the one type that is not considered symmetrical. That is, while the others go in both directions, with projection we are predicting the

[2] Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*

score on one test based on that of the other. That prediction does not work in the reverse direction. *Concordance* is what we see when there are two tests that fulfill a similar purpose but are designed to different frameworks and specifications. For example, studies have been done to map the scores of the SAT to the ACT, or NAEP to TIMSS.

The three level tests of either ECLS-K Reading or ECLS-B Reading can be considered to satisfy the requirements of *calibration* and one method that is often associated with that linking type is called *concurrent calibration* and is explained further below. Across cohorts, the link is not as strong for a number of reasons, the most important being the time lapse and the somewhat different framework. For this reason, I am performing a linking that satisfies the claims of *statistical moderation* for that cross-cohort situation.

Selection of Linking Method

Below are some of the basic types of linking methods available to us:

1. Identity
2. Mean
3. Linear equating
4. Equipercentile
5. IRT Concurrent Calibration
6. IRT Test Characteristic Curve
7. IRT Mean & Sigma

The first four are Classical Test Theory methods and have the limitation that they are sample dependent. *Identity* linking is the null case. One simply decides, for example, that a 50 on test 2 means the same as a 50 on test 1. It would require that both tests have the same number of items. With *mean equating* or linking, one obtains the means of the scores for the two tests, finds the difference and then adds (or subtracts) that difference to the focal test scores. *Linear equating* takes this one step further by taking the standard deviation into account and applying a constant and a slope to the scores. For the e*quipercentile* method the scores of students at a given percentile on the reference test are matched to those on the focal test. In this way a table of score equivalents is constructed.

The IRT methods are often preferred because the scale scores resulting from an IRT analysis are considered to be sample invariant, unlike methods 1 through 4 in the list above. With *concurrent calibration*, item level scores from tests that have some items in common and some items that are unique are processed all together in order to produce estimates of item difficulty or to produce examinee scores. This is possible because IRT calculations are robust to missing values. The *IRT Test Characteristic Curve* method (Stocking & Lord, 1982) involves mapping scores from one test onto another according to the probability ogives known as the Test Characteristic Curves. *IRT Mean and Sigma* (Marco, 1977) methodology can be used for chain linking and fixed parameter methods of re-scoring. In this method, constants for the IRT parameters of the reference test are identified and applied to the focal test scores in order to produce focal test scores that are on the same scale as the reference test. IRT *Mean and Sigma Equating* is what I have used in linking ECLS-K and ECLS-B in the present analysis as it is fairly straightforward and can be performed without the use of any specialized software outside of PARSCALE and MS Excel.

Scores Summaries Before and After

After linking and rescaling, new scores were created for all participants in the ECLS-K and ECLS-B who were between the ages of 57-75 months and who were taking the test as first-time kindergartners. Table 3 shows the difference in the old and new score distributions for these children.

Table 3. Scores Summaries Before and After Linking and Rescaling

	Original Scores				
	n	Mean	St. Dev	Min	Max
ECLS-K	3,341,700	-1.33	0.51	-2.71	0.92
ECLS-B	2,787,700	0.49	1.02	-9.00	3.09
Difference		***1.82***			
	New (Linked) Scores				
	n	Mean	St. Dev	Min	Max
ECLS-K	3,341,700	-0.80	0.82	-3.14	2.98
ECLS-B	2,787,700	-0.17	0.86	-2.52	2.57
Difference		***0.63***			

NOTE: These descriptive statistics were produced using the complex sample statistical software package AM. ECLS-K 1998 estimates centered on overall weight C1CW0: C1 CHILD WEIGHT FULL SAMPLE , while ECLS-B 2006 estimates centered on overall weight WC40: W4 RSP1/RSP2/RSP3/RSP4/C1/C2/C3/C4-FULL SMP WGT.
SOURCE: Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort, Kindergarten-Year Reading Assessment public use data set (NCES 2010-010). National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Cohort, Kindergarten-Year Reading Assessment public use data set (NCES 2009-005).

Recall that the original scores are not on the same scale even though they both use a theta scale. With that caveat firmly in mind, we see that for the means of ECLS-K (using the theta metric that usually ranges from approximately -4 to +4) the original scores were dissimilar across the two studies. But we cannot be sure how dissimilar they were because while they were both produced using IRT calibration, they were not on the same theta scale. That is, they were not linked. I made them more comparable in a theoretically justifiable way during my analysis by reducing the tests to only items that were more directly comparable in terms of the general reading construct being measured. A well-fitting IRT model was applied to the reference test after it was reduced to only the item set that represented one general reading construct. The resulting item parameters were adjusted using the IRT mean and sigma approach and then applied as fixed values to the common items in a new calibration of the focal test, ECLS-B Reading. An output of the new calibration of the ECLS-B test was a new set of scores. Once these adjustments were made, the refined assessments across the studies could be said to have comparable scores. Looking at table 3, we see that, in the original scores, the ECLS-B students are 1.82 theta points above the ECLS-K. In the new scores, the ECLS-B students are .63 theta points above the ECLS-K, about a third as much of a difference as in the non-linked score comparison. In addition, we see that the standard deviations, minimum scores and maximum scores of both sets of scores are more in line with each other.

Discussion of Cross Cohort Linking and Vertical Scaling

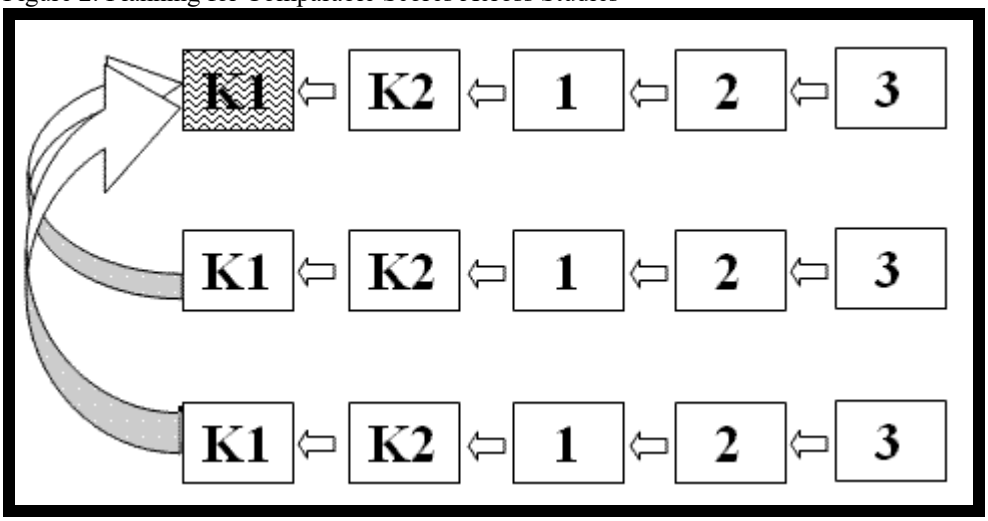
At this point I have demonstrated that there is at least one way to link scores from the tests of two disparate cohorts in the ECLS. As a reminder, here was the process employed for this analysis:

Step 1. Create restructured item level data sets for ECLS-K and ECLS-B kindergarten reading.
Step 2. Perform factor analysis on the level test item sets. Remove items that do not align with primary factor efficiently across level tests.
Step 3. Determine if tests are similar enough for comparison based on criteria discussed by Kolen and Brennan (2004).
Step 4. Identify the best fitting IRT model for both ECLS-K and ECLS-B kindergarten reading tests, estimate items parameters for both, continue to assess comparability, and obtain new scores for ECLS-K.

Step 5. Apply item parameters of the ECLS-K test (selected common item set noted in step 2) to the matching items of the focal test (ECLS-B) and calibrate with all other items of ECLS-B, and create a new set of parameters and scores for ECLS-B that are on the same scale as ECLS-K.

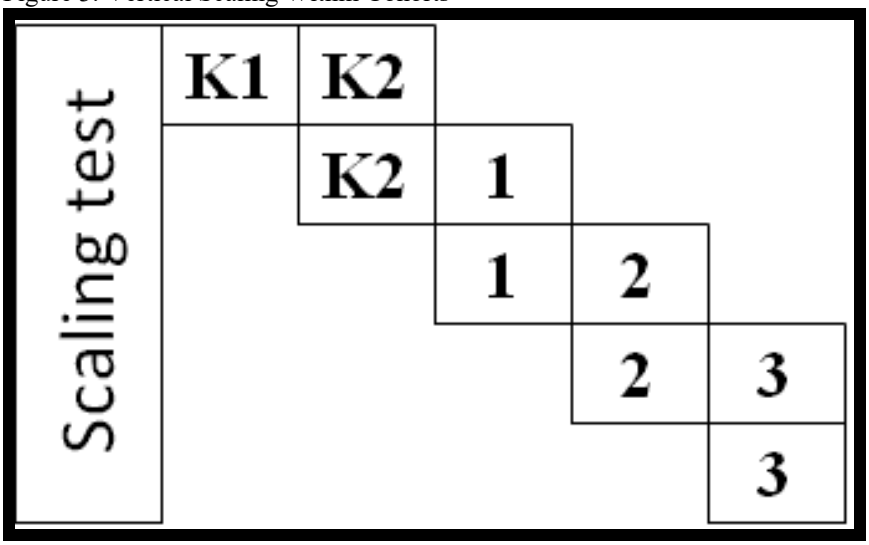
By establishing that a linking of these two tests can be done, we are free to move on to consider how we can plan in the long term for linking across longitudinal studies. In an ideal situation, the framework for the domain is stable over time and the new forms for each cohort are developed in accordance with that stable framework. Figure 2 shows a scenario in which the base year ECLS-K provides the reference for subsequent ECLS kindergarten tests. Those tests, if designed to exactly the same framework and test specification, could even be treated as parallel forms and thus meet the requirements of score *equating*, rather than the looser *statistical moderation*. Even if equating is not practical, the link could go forward, with each subsequent kindergarten test serving as the referent for the vertical scale within the cohort.

Figure 2. Planning for Comparable Scores Across Studies



Within cohorts, items that overlap from one time point to the next are used to create the vertical link. This is illustrated in figure 3. For vertical scaling, all test development is carried out after a grade-to-grade (or round-to-round) model of growth in the domain has been defined, reviewed qualitatively by subject matter experts, and prototyped in item exemplars or assembled in a draft item pool. In theory, there is a scaling test that tests the full range of the content through all developmental levels. The full scaling test is never administered in practice. There would not be adequate variability for the items at all levels, nor would there be enough time to take such a test. In practice, we develop the vertical scale by administering forms to examinees in a calibration sample at each grade level. The item content overlaps with the next level up so that a score link can be made.

Figure 3. Vertical Scaling Within Cohorts



Limitations of this Study

This study remains in the domain of exploration and possibility. The goal is to inform process and design of the academic tests of longitudinal cohort studies at NCES and elsewhere. Limited resources were devoted to this study and as such there were some analysis types that could not be attempted. For the data itself, the common item sets were not a mirror image of the entire test. They tended to be the lower level items of the stage 1 test. In planning for the instruments of future tests, we would want to attempt to embed a "mini-test" throughout the instrument, one with items in a variety of subdomains and difficulty levels. Another weakness of the link is an indeterminate unidimensionality across levels or cohorts. This could only be checked qualitatively. With the use of more sophisticated software such as TESTFACT (SSI, Inc. 2003) we may be able to confirm this characteristic of the tests across cohorts in an empirical fashion. With other IRT software such as IRTEQ (Han, 2008) we might compare other linking methodologies such as Stocking and Lord (1983) test characteristic curve to the means and sigma approach. By deriving a standard error of equating we can select among methods for the most precise.

Future Directions

Once there is confidence in the comparability of ECLS-K and ECLS-B kindergarten reading scores, one can consider some of the follow-on activities possible. NCES may consider publishing an additional set of scores for ECLS-K round 1 and ECLS-B round 4 to be used for comparability studies. With linked score data, researchers could investigate how certain covariates perform in the two studies relatively speaking. One could even investigate applying these principles in a vertical scaling revision of the ECLS-K Round 2-7 scores. As NCES moves forward with any new longitudinal studies, new tests can be designed with an eye toward preserving a well-defined and validated set of linking items. Our results in subsetting the data indicate that funds could be used more efficiently or effectively by using fewer examinees per round of field testing in the test development phase than in the past and thus permit a more iterative item selection and test development process or a less costly single field test.

References

AM Statistical Software (v.0.06.00). American Institutes of Research http://am.air.org/. Last accessed on 11/15/2011.

Brennan, R. L., National Council on Measurement in Education, & American Council on Education. (2006). *Educational measurement* (4th ed.) Westport, CT: Praeger Publishers.

Department of Education, National Center for Education Statistics. ECLS-K website, http://nces.ed.gov/ecls/kindergarten.asp.

Department of Education, National Center for Education Statistics. (2010). Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K) Correct Theta Scores for the Kindergarten through Eighth-Grade Data Collections Errata. (NCES 2010-052). Washington, DC.

Department of Education, National Center for Education Statistics. (2009). Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) 9-Month—Kindergarten 2007 Restricted-Use Data File and Electronic Codebook (CD-ROM). (NCES 2010-010). Washington, DC.

Department of Education, National Center for Education Statistics. (2009). Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K) Eighth Grade Restricted-Use Data File and Electronic Codebook (DVD). (NCES 2009-006). Washington, DC.

Department of Education, National Center for Education Statistics. (2009). Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K) Kindergarten through Eighth Grade Full Sample Public-Use Data and Documentation (DVD). (NCES 2009-005). Washington, DC.

Department of Education, National Center for Education Statistics. (2009). Early Childhood Longitudinal Study, Kindergarten Class of 1998 - 99 (ECLS-K), Psychometric Report for the Eighth Grade NCES Number: 2009002 Release Date: September 25, 2009

Du Toit, M. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International, Inc.

Fidelman, C. G. (forthcoming). Linking of results for ECLS-K Round 1 and ECLS-B Round 4 Kindergarten Math and Reading Assessments . NCES Publication, in development.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, Calif.: Sage Publications.

Han, K. T. (2009). IRTEQ: Windows application that implements IRT scaling and equating [computer program]. *Applied Psychological Measurement*, 33(6), 491-493.

Kolen, M. J., & Tong, Y. (2009). Vertical scaling training session. American Education Research Association Annual Meeting, San Diego.

Kolen, M. J., Brennan, R. L., & Kolen, M. J. (2004). *Test equating, scaling, and linking : Methods and practices* (2nd ed.). New York: Springer.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, 14(2), 139-160

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17*(2), 131-154. doi:10.2307/1165166

Muraki, E., & Bock, R.D. (2003). PARSCALE 4.1 for Windows: IRT based test scoring and item analysis. Chicago: Scientific Software International, Inc.

Rao, C. R., & Sinharay, S. (2007). Psychometrics (1st ed.). Amsterdam ; Boston: Elsevier North-Holland.

Stocking, M. L., Lord, F. M., & Educational, T. S. (1982). *Developing a common metric in item response theory*. Princeton, NJ: Educational Testing Service.

Wang, S., & Jiao, H. (2009). Construct equivalence across grades in a vertical scale for a K-12 large-scale reading assessment. *Educational and Psychological Measurement, 69*(5), 760-777.