Summary of technical information

Heather C. Hill

02/04

This document provides a guide to existing forms and scaling information. It is meant to help potential measures users determine whether our pre-piloted forms/scales will be sufficient, technically speaking, for their needs. Potential users should keep in mind that they may also create their own forms from our item pool, and thus may improve the reliability and targeting of the scales.

In the first two sections, we provide a brief history of our project and outline the forms/constructs that have been piloted to date. In the third section, we provide tables with psychometric information for each piloted scale. Finally, we answer some general questions about the items.

History of project

The Study of Instructional Improvement (SII) began writing and piloting items on a large scale in 2001. With the cooperation of the University of California Office of the President, three forms were piloted in Mathematics Professional Development Institutes in summer 2001. Using both items from these three forms and new items written in 2002, another two forms were piloted in 2002-2003 MPDIs. In 2002, we also piloted geometry items with the Learning Mathematics for Teaching (LMT) "comparison group." Finally, we have built new forms for 2004, based on requests from MSPs and others for coverage of particular content domains, and on results from 2002 piloting.

In 2004-2005, we have also begun to pilot middle school forms containing number/operations, pre-algebra and algebra, and the set of geometry items described above.

We refer to forms by their year of pilot (2001, 2002, or 2004) and identifying letter (A,B and C). Forms were built to be parallel to one another (equated) within year of pilot (i.e., A,B and C from 2001 can be equated, but A01 and A02 cannot).

Scales and forms

Each form carries a unique set of items. For instance, forms A, B and C in 2001 were designed to be parallel (equated) to one another. They carry common linking items, but the bulk of items are different across the three forms. This enables pre-post comparisons without significant concern about test-retest effects. Also, some constructs (e.g., patterns, functions, and algebra) were piloted some years but not others.

To provide a record of what was piloted when, we list the scales piloted on each form below. Users interested in pre-piloted scales may mix and match scales from various

piloted forms, with the constraint that there is no cross-year equating for scales, and that 2004 forms have not been equated yet.

Before choosing to use any particular set of forms, users should check the scaling properties presented below and in the technical reports in this CD.

Here is a guide to scales and forms:

2001 pilot. Forms A01, B01 and C01 contain the following scales:

- Number and operations content knowledge
- Patterns, functions and algebra content knowledge
- Number and operations knowledge of students and content

2002-2003 pilot. Forms A02 and B02 contain the following scales:

- Number and operations content knowledge
- Number and operations knowledge of students and content

LMT comparison group pilot. This comparison group form (CG03), administered to teachers not participating in the MPDI, was used to pilot geometry items. It also includes other content areas. The three constructs assessed on this form are:

- Number and operations content knowledge (same as Form B02)
- Geometry content knowledge, grades 3-8
- Patterns, functions, and algebra content knowledge (collected items from 01 piloting)

The geometry items may not be an ideal fit for K-2 teachers.

2004 elementary forms (new). Forms MSP_A04 and MSP_B04 contain items to comprise the following scales:

- Number and operations content knowledge
- Patterns, functions and algebra content knowledge
- Geometry content knowledge, grades 3-8

The number and operations CK scales are based on those on the 2002-2003 pilot. They have, however, had items added to improve measurement. In addition, two items were removed from Form B02: B8 and B18D.

The patterns, functions, and algebra CK scales are based on Form A and Form C from the 2001 pilot. These too have had items added to improve measurement.

The geometry items were taken from the comparison group pilot. No new items were added. They have acceptable measurement and targeting.

2004 middle school forms (new). Forms Middle-A and Middle-B contain items to comprise the following scales:

- Number and operations content knowledge, 6-8
- Patterns, functions and algebra content knowledge, 6-8
- Geometry content knowledge, grades 3-8

The geometry scales are identical to the ones included on the elementary forms, with the exception of one additional item per form. In the other two content areas, most items are new, although a small number of difficult elementary items allow linking the forms across populations. No piloting information is available as of 12/04.

Additional sections and questions. Each form has additional questions in Section 1. Some of these questions ask about teachers' classroom teaching, experiences, and background. We have used these questions to explore whether teachers' classroom teaching practices, years of experience in the classroom, or grade level relate to measure score. You need not use these items, but should note that grade level is often related to measure score – you may wish to control for this in your own analyses.

Some forms also have teacher motivation questions. You might use these to explore whether teachers who are more motivated to learn do so.

Finally, post-test forms have a set of questions asking about the perceived content of professional development. We have used these to make comparisons among different kinds of professional development experiences for teachers – to ask whether gains in teachers' content knowledge is related to working on artifacts from classroom practice, to working with other teachers, to proving and solving lots of mathematics problems for themselves. If you choose to have teachers participate by completing this section of the form, it means that data from your site may be used in the future for such comparisons.

Summary of psychometric information about forms

Below we have reproduced basic information about scales for each form, along with some comments about scales on each form. These are organized by scale, rather than form, for ease of comparison. More detail is presented in the technical reports for each scale.

In examining the reliabilities below, it is helpful to reference some established rules of thumb. Reliabilities of .7 are generally adequate for finding moderate effects in groups of 60 or more individuals. Reliabilities of .8 are generally adequate for finding small effects (.3 standard deviation) in groups of 60 or more. And reliabilities of .9 or higher are necessary for making claims about individuals' (as opposed to group's) levels of knowledge.

As these guidelines suggest, results from pre-piloted forms are not reliable enough to make statements about individual teachers' levels of knowledge – for instance, saying Teacher A has inadequate knowledge, or Teacher B has more knowledge than Teacher A. Another way to think about this is to notice that the error associated with any individual is sufficiently high to prevent making confident statements about their ability. Instead, we use these items to make statements about how content knowledge differs among groups of teachers, or how a group of teachers performs at one or more time points. Usually, with sufficient numbers of teachers per group (N=60 or more) errors in measurement are of less concern with the reliabilities we report below.

There have been six forms (A01, B01, C01, A02 B02, Comparison Group03) piloted to date. Not every measure was contained on every form. We organize our description by measure, rather than form.

Content knowledge items in number/operation. Overall, reliabilities were adequate (.70 or above) for most content knowledge measures piloted in 2001-2003. In 2001, reliabilities were high (.72-.80) but scales were poorly targeted – the best measurement occurred for individuals one-half to one standard deviation below average. In 2002-2003 piloting, targeting was improved but reliabilities were lower. Lower reliabilities reflect, in part, shorter forms (e.g., n items = 16 on form B02). There are also possible ceiling effects in all piloted scales, making these measures potentially inappropriate for detecting growth among already-highly-performing teachers. We attempted, with the 2004 form, to add items to the 2002 forms to increase reliability and ameliorate ceiling effects.

Table 1: Two-parameter model, all data all items

Form	Sample size	Number of items	2 PL reliability*	Max information (2PL)	Number of items over 1SD (2PL)
A01	652	26	.80	-.25	4
B01	599	25	.83	-1.13	1
C01	377	23	.83	-.25	2
A02	1020	24	.82	-.37**	2
B02	1095	16	.75	-.37	1

* Reliabilities vary less than two hundredth of a point from that presented in Hill, Schilling & Ball (in press). This discrepancy arises because we used the newer version of Bilog (MG) for model estimation in 2004, when this report was written.
**Targeting for LMT comparison group sample; form A used as post-test in larger MPDI sample, and thus targeting is off.

Table 2: One-parameter model, pre-test only

	Sample size	Number of items	reliability	Max information	Number of items over 1SD
A01	411	26	.72	-.75	1
B01	159	25	.80	-1.0	1

C01	104	23	.76	-.5	2
A02*	454	24	.81	-.37	2
B02	1095	16	.72	-.25	1

*Estimated on the LMT comparison group

Content knowledge in patterns, functions, and algebra. These items scale well (.70 with 12-15 items). However, they are poorly targeted. Ceiling effects are likely with the 2001 scales. In 2004, we added items we anticipate are more difficult, in an attempt to improve measurement.

Table 3: Two-parameter model, all data all items

	Sample size	Number of items	2 PL reliability	Max information	Number of items over 1SD
A01	652	12	.77	-.75	0
B01	599	12	.78	-1.25	0
C01	377	10	.73	-1.12	0

* Reliabilities vary less than two hundredth of a point from that presented in Hill, Schilling & Ball (in press). This discrepancy arises because we used the newer version of Bilog (MG) for model estimation in 2004, when this report was written.

Table 4: One-parameter model, pre-test only

	Sample size	Number of items	1 PL reliability	Max information	Number of items over 1SD
A01	411	12	.65	-.5	0
B01	159	12	.74	-1.12	0
C01	104	10	.65	-1.25	0

Forms A01 and C01 were used as the basis for forms A04 and B04, respectively. Items were added to improve reliability and targeting.

Content knowledge in geometry. All geometry items (n=43) were piloted on one form, the comparison group post-test 2003. The maximum reliability of this set of items is .95. Parallel forms (A and B) with roughly 25 items have reliabilities of roughly .85.

Knowledge of students and content. Generally, reliabilities here are not sufficient to detect moderate effects in mid-sized groups. We recommend construction of new scales and forms from the item pool.

Table 5: Number and operation KSC, two-parameter model, all data all items

	Sample size	Number of items	2 PL reliability	Max information (2PL)	Number of items over 1SD (2PL)

A01	652	20	.70	-1.75	1
B01	599	18	.73	-1.0	0
C01	377	21	.77	-1.0	2
A02**	1020	16	.57	-1.4	2
B02	1095	16	.73	-1.5	1

* Reliabilities vary less than two hundredth of a point from that presented in Hill, Schilling & Ball (in press). This discrepancy arises because we used the newer version of Bilog (MG) for model estimation in 2004, when this report was written.
** This form was used solely as a post-test for the MPDIs. Item difficulties are likely underestimated. Unfortunately, there is no comparison group data for estimation of more correct difficulties.

Table 6: Number and operation KSC, one-parameter model, pre-test only

	Sample size	Number of items	reliability	Max information	Number of items over 1SD
A01	411	20	.58	-1.6	1
B01	159	18	.67	-1	1
C01	104	21	.58	-1.4	3
A02**	1020	16	.44	-1.5	3
B02	1095	16	.66	-1.12	0

** This form was used solely as a post-test for the MPDIs. Item difficulties are likely underestimated. Unfortunately, there is no comparison group data for estimation of more correct difficulties.

A note about ceiling effects

Ceiling effects occur with some of the scales and samples reported here. This makes our pre-piloted measures unsuitable for evaluations of programs where teachers enter with high levels of mathematical knowledge. This also means the K-5 pre-piloted forms cannot be used with middle school teachers, for the most part.

Users interested in evaluating programs of this type may elect to build their own forms from more the more difficult items in our pool.

A note about content coverage

One issue in test design is the sampling of problems from the domain of all problems that theoretically exist in a given area. The SII/LMT measures developers elected to sample problems for mathematics teaching that would yield the best discrimination among teachers, rather than writing problems to represent K-5 mathematics teaching content broadly. For this reason, these items do not perfectly match topics covered in K-5 mathematics – we are missing items in the domain of early-grade addition and subtraction, for instance. Users should refer to the project's content validity documents for a report on areas included and not included in our item pool.

A related issue is how content is represented in a particular test form. In general, we believe content was represented most broadly by the 2001 piloted forms. In 2002-2003, we selected a smaller set of items for forms A and B. On form B, in particular, there are only 10 stems – i.e., 10 problems sampled from the domain of elementary mathematics. While we are able to generate reasonably reliable estimates of individuals' knowledge from these 10 stems (and 16 items), this form may not be particularly sensitive to professional development programs, since its representation of content is so very selective.

Status of equating

Elementary umber and operations content knowledge scales have been equated for the 2001 and 2002 years. The equating seems sufficient. Our 2004 elementary number and operations scales have not been equated, but users may revert to the 2002 equating (2004 forms are built from 2002 forms).

Number and operations knowledge of students and content scales have been equated for 2001. We have concerns about the accuracy of this equating, yet no plans to conduct further analyses at this time. The 2002 forms have not yet been equated.

Geometry content knowledge items have been equated, and we are highly confident in the results.

Middle school number/operations and algebra items have not been equated as of spring 2005.