ICPSR | INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH

ICPSR 36095

National Center for Teacher Effectiveness Main Study

Thomas Kane
Harvard Graduate School of Education

Heather Hill
Harvard Graduate School of Education

Douglas Staiger
Dartmouth College

Construct Technical Report

Inter-university Consortium for
Political and Social Research
P.O. Box 1248
Ann Arbor, Michigan 48106
www.icpsr.umich.edu

Terms of Use

The terms of use for this study can be found at:
http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/36095/terms

Information about Copyrighted Content

Some instruments administered as part of this study may contain in whole or substantially in part contents from copyrighted instruments. Reproductions of the instruments are provided as documentation for the analysis of the data associated with this collection. Restrictions on "fair use" apply to all copyrighted content. More information about the reproduction of copyrighted works by educators and librarians is available from the United States Copyright Office.

Construct Technical Reports

Table of Contents

Section 1. Teacher Questionnaire

Section 1.a.i. Fall Questionnaire - Overview

NCTE administered surveys to participating teachers in the fall of each data-collection year. In addition to items assessing teachers' knowledge, as determined by their performance on *Mathematical Knowledge for Teaching* (MKT) and *Massachusetts Tests for Educator Licensure* (MTEL) items, the fall questionnaires contained a series of items probing teachers' instructional practices, beliefs about teaching and learning, experiences with professional development and evaluation, and perceptions of their school environments. Most of the items on the fall questionnaire remained constant across the three years of data collection; however, a few item sets were revised, retired, or added in later years.

For the questionnaire, exploratory factor analysis (EFA) of the teachers' responses to the non-MKT/MTEL items was used to determine the factor structure of each set of survey items designed to measure a latent construct. The results of the EFA informed the creation of composite scales to be used in analysis, as well as decisions about revising, deleting, or adding items to the following year's survey.

In year one of NCTE, 249 of 263 surveys sent out were completed, for a response rate of 95%. In year two, this rate was 98% (214/219), and in year three, this rate was 96% (178/186). In what follows, we describe the technical aspects of the construction of various constructs captured in the questionnaire.

Section 1.a.ii. Fall Questionnaire – Teacher Knowledge (MKT/MTEL)

We surveyed teachers in the fall of each of three years. One section of that survey consisted of mathematics items designed to measure teachers' mathematical knowledge. Some of those items were taken from the Mathematical Knowledge for Teaching (MKT) item bank, and others were released items from the Massachusetts Tests for Educator Licensure (MTEL). The idea was to measure two types of mathematical knowledge for teaching—"specialized content knowledge", theorized to be represented by the MKT items, and "common content knowledge", theorized to be represented by the MTEL items. In the end, we were unable to support a 2-factor structure from the data, and produced a single "teacher knowledge" score for each teacher.

In Year 1, there were 32 MKT items and eight MTEL items.[1] We received responses from 247 teachers. In Year 2, there were 24 MKT items and 12 MTEL items; we received responses from 214 teachers (i.e., N = 301 teachers across Years 1 and 2). The Year 3 form had 16 MKT items and 13 MTEL items; we received responses from 176 teachers (i.e., N = 313 unique teachers across all years).

If six or more consecutive items were missing, we chose to score those items as "not presented" rather than incorrect.[2] Any missing response not within a block of six or more missing responses was scored as incorrect.

Before we conducted exploratory factor analyses (EFAs) on the entire set of items to look for different types of knowledge constructs, we explored responses to the "testlets" (i.e., instances where multiple items were nested under a single question stem) embedded within our surveys. Responses to these nested items are likely to be correlated, which will be problematic for factor analyses, and also for any scoring paradigm that assumes the independence of responses, conditional on ability.

To investigate testlet response for each year, we first treated each item as a separate response, and created a tetrachoric correlation matrix using all teachers who had responses to every item (Year 1 N = 235; Year 2 N = 204; Year 3 N = 174). In Year 1, there were five testlets among the 32 items. A 5-factor EFA (see Figure 1) showed that for four of those testlets, the subitems were strongly intercorrelated. The final testlet comprised five unrelated true/false geometry items. These results were the same when forcing the existence of four or six factors.

Unfortunately, the Years 2 (see Figures 2 and 3) and 3 (see Figures 4 and 5) data were not nearly as clear. Of the four nominal testlets in Year 2, one had items strongly load on a factor, another had items weakly load on a factor, and items from the other two were (weakly) split across factors. In Year 3, the three nominal testlets were all split across factors. These results tended to hold even as we changed the number of factors forced. Unlike the final testlet from Year 1, which comprised unrelated true/false questions, all seven of these Years 2 and 3 testlets were substantively linked.

---

[1] Though the items on the forms were not marked as coming from different sources, in Year 1 the MTEL items appeared at the end of the form, after all of the MKT items. In Years 2 and 3, the MTEL and MKT items were intermixed.

[2] The inference being that teachers skimmed over these items rather than read them and then chose not to answer them.

Given the ambiguity of the empirical results, and the relatively moderate sample size of teachers, we relied on theory and rescored the items to reflect the presence of 11 testlets—all of the nominal testlets, except for the final testlet comprising true/false items in Year 1. To do so, we summed the correct number of subitems for each testlet to create a single, ordinal score.

Next, we dropped poorly performing items in preparation for the overarching EFA. We first created six "forms" of responses: all Year 1 items, all Year 2 items, all Year 3 items, all MKT items across years, all MTEL items across years, all items across years. Note that each item appears in exactly four of these forms. If an item had an item-rest correlation of .15 or lower on three out of the four forms, it was dropped. None of the ordinal-scored testlets were problematic, so we went back and used the data from scoring all items independently to search for poorly performing testlet subitems. We also ran 2-parameter IRT models on all of the different forms, and investigated extreme slope and discrimination values. 12 items were dropped as a result of these processes: two MKT items from Year 1, four MKT items and 2 MTEL items from Year 2, and two MKT and one MTEL item from Year 3. A 13^{th} item (MTEL from Year 2) was dropped because 98% of teachers answered it correctly, which resulted in problems with some polychloric correlation convergences.

Once these items were removed, we explored the factor structure of the data. EFAs on Years 1 and 3 data supported a 1-factor solution, with a 2-factor solution not well supported. EFAs on Year 2 data supported a 2-factor solution or possibly a 1-factor solution. Running an EFA on all items pooled across years supported a 1-factor solution. Since we had substantive reason to believe the data represented two factors, we forced 2-factor solutions and examined the factor loadings (see Figures 6, 7, 8, 9, 10, and 11). Item analysis did not reveal any obvious themes or similarities amongst items within proposed factors.

As a further validity check, we asked four math education experts to blindly categorize items as either specialized content knowledge or common content knowledge. Combining their rankings into bins of 'common', 'mixed', or 'special', the experts' categorizations largely matched the theorized categorization by item source; that is, all but one common item were MTEL items, and all but one specialized item were MKT items (see Figure 12 below). Re-examining the EFAs above in terms of these item characterizations did not change any of the interpretations.

Given that we could not find reliable empirical support for our theorized 2-factor structure, we treated the data as unidimensional. We then scored the data as a single form with all items across all years (except for the 13 dropped items, and with testlets scored as ordinals). We used a 1-parameter graded response model, scored in IRTPRO.

Variable	Factor1	Factor2	Factor3	Factor4	Factor5	Uniqueness
mkt01y1	0.2102	0.1	0.3326	0.528	0.0114	0.5562
mkt02y1	0.1371	0.0777	0.0504	0.7799	0.1547	0.3405
mkt03y1	0.2762	0.0841	0.0695	0.4702	0.5351	0.4045
mkt04y1	0.2526	0.1869	0.1896	0.8417	-0.0167	0.1566
mkt05y1	0.5702	0.1988	0.463	0.0196	0.2569	0.3546
mkt06y1	0.2152	0.3419	0.5625	0.1312	-0.1991	0.4636
mkt07y1	0.2189	0.3311	0.3486	0.026	0.494	0.4763
mkt08y1	0.6425	0.0096	0.3627	0.1202	0.0984	0.4314
mkt09y1	0.453	0.2833	0.3487	0.0116	0.1368	0.5741
mkt10y1	0.4522	0.2024	0.2996	-0.2655	0.1468	0.5727
mkt11y1	0.176	0.0771	0.2359	0.3766	0.4361	0.5755
mkt13y1	0.0907	0.3485	0.3519	-0.1733	0.4067	0.551
mkt14y1	0.3532	0.2045	0.0048	0.0858	0.4881	0.5879
mkt15y1	0.4155	0.4267	0.2243	0.0867	0.3658	0.4536
mkt16y1	0.4619	0.1769	0.4611	-0.1866	0.1222	0.4929
mkt17y1	0.1356	0.1568	0.8144	0.1499	0.1586	0.2462
mkt18y1	-0.027	-0.008	0.7802	0.2167	0.0545	0.3405
mkt19y1	0.0738	0.1727	0.4704	0.2219	0.2543	0.6295
mkt21y1	0.3014	0.3235	0.2348	0.062	0.2864	0.6635
mkt22y1	0.0218	0.3797	0.0991	0.0717	0.6061	0.4731
mkt23y1	0.0145	0.734	-0.021	0.2487	0.2774	0.3218
mkt24y1	0.0782	0.9009	0.226	-0.0231	0.0467	0.1285
mkt25y1	0.2591	0.7509	-0.0466	0.1428	0.015	0.3462
mkt26y1	0.4444	0.529	0.0426	0.0548	-0.1718	0.4883
mkt27y1	0.1456	0.9265	0.0933	0.0992	0.1222	0.0868
mkt28y1	0.8601	0.1764	-0.0688	0.0951	0.2732	0.1408
mkt29y1	0.6787	0.047	0.1432	0.1959	-0.2141	0.4325
mkt30y1	0.6782	0.0569	0.0765	0.2433	0.0349	0.4705
mkt31y1	0.8047	0.2508	0.0852	0.3647	0.1251	0.1337
mkt32y1	0.387	0.048	0.16	0.0712	0.5923	0.4665

Figure 1. Year 1 testlet factor loading results. Colored variable names indicate testlet groupings

Variable	Factor1	Factor2	Factor3	Factor4	Factor5	Uniqueness
mkt01y2	-0.0003	0.1178	0.0132	0.0357	0.9183	0.1414
mkt02y2	0.4179	-0.1895	-0.2616	0.4623	0.2802	0.4287
mkt03y2	0.2528	0.0434	0.6619	0.1848	-0.2114	0.4172
mkt04y2	0.0438	0.025	0.7452	-0.0506	0.097	0.4302
mkt05y2	-0.0184	0.1507	-0.1538	0.601	-0.0497	0.5896
mkt06y2	0.2337	0.6082	-0.0938	0.0682	0.3836	0.4149
mkt07y2	0.2484	0.6318	0.1245	-0.1525	0.5231	0.2268
mkt09y2	0.767	0.4964	-0.0656	0.1349	0.1958	0.1044
mkt10y2	0.2866	0.026	0.2776	0.5623	0.2019	0.4833
mkt11y2	0.4311	0.3634	0.3224	-0.0609	0.0265	0.5737
mkt12y2	-0.0692	0.2098	0.1821	0.6596	-0.2964	0.3951
mkt13y2	0.7335	-0.127	0.2709	0.0086	0.3938	0.2172
mkt14y2	0.1535	0.4151	0.5106	0.0824	0.3919	0.3831
mkt15y2	0.4525	0.2117	0.5112	0.2533	0.0041	0.425
mkt16y2	0.8241	0.062	0.1326	-0.0327	-0.3388	0.1835
mkt17y2	0.4356	0.2905	0.4082	0.3789	0.1666	0.3879
mkt19y2	0.1452	0.3242	0.3213	0.376	0.0504	0.6267
mkt21y2	-0.1121	-0.255	0.4075	0.4801	0.3456	0.4064
mkt22y2	0.3996	0.38	0.0659	0.4822	0.2735	0.3843
mkt24y2	0.0054	0.802	0.1213	0.1942	-0.092	0.2959

Figure 2. Year 2 testlet factor loading results, 5-factor solution. Colored variable names indicate testlet groupings

Variable	Factor1	Factor2	Factor3	Factor4	Uniqueness
mkt01y2	0.2817	0.0315	-0.0983	0.8759	0.1427
mkt02y2	0.1865	0.0931	0.2218	0.2175	0.86
mkt03y2	0.0174	0.6749	0.2308	-0.1945	0.4532
mkt04y2	-0.0585	0.6335	0.0019	0.1264	0.5793
mkt05y2	0.097	-0.1139	0.5932	0.0219	0.6253
mkt06y2	0.6958	-0.0651	0.1015	0.279	0.4235
mkt07y2	0.7235	0.1034	-0.1039	0.3922	0.3012
mkt09y2	0.8754	0.2505	0.066	-0.0171	0.1663
mkt10y2	0.1626	0.4327	0.4596	0.2171	0.528
mkt11y2	0.4831	0.4032	-0.0193	-0.087	0.5961
mkt12y2	0.0009	0.1199	0.7473	-0.1716	0.3978
mkt13y2	0.3702	0.6427	-0.2136	0.2184	0.3565
mkt14y2	0.4007	0.4377	0.1326	0.3621	0.4992
mkt15y2	0.3314	0.6318	0.2476	-0.0489	0.4273
mkt16y2	0.4363	0.4958	-0.1015	-0.5285	0.2742
mkt17y2	0.429	0.5406	0.3494	0.1164	0.3881
mkt19y2	0.2747	0.3084	0.4248	0.0702	0.644
mkt21y2	-0.2734	0.397	0.3731	0.4665	0.4108
mkt22y2	0.5528	0.2345	0.4252	0.2128	0.4133
mkt24y2	0.5677	-0.0544	0.4176	-0.0943	0.4914

Figure 3. Year 2 testlet factor loading results, 4-factor solution. Colored variable names indicate testlet groupings

Variable	Factor1	Factor2	Factor3	Factor4	Uniqueness
mkt01y3	0.5174	0.1801	0.4898	-0.3315	0.35
mkt02y3	0.7551	-0.0267	-0.051	0.0914	0.4181
mkt04y3	0.8321	0.0179	-0.0291	0.1599	0.281
mkt06y3	-0.0263	0.7991	0.1887	-0.2501	0.2625
mkt08y3	-0.0163	-0.0169	0.7504	0.2483	0.3747
mkt09y3	0.1365	0.0232	0.2145	0.8763	0.167
mkt10y3	-0.007	0.0128	0.7026	0.2575	0.4399
mkt11y3	0.0328	0.7013	0.0169	0.2071	0.464
mkt12y3	0.3154	0.3105	0.442	-0.0033	0.6088
mkt14y3	0.3996	0.6601	-0.2622	0.2157	0.2893
mkt15y3	0.5904	0.4799	0.05	0.1203	0.4042
mkt16y3	0.6559	0.2552	0.3149	-0.0376	0.4041

Figure 4. Year 3 testlet factor loading results, 4-factor solution. Colored variable names indicate testlet groupings

Variable	Factor1	Factor2	Factor3	Uniqueness
mkt01y3	0.4286	0.2919	0.3292	0.6227
mkt02y3	0.7586	-0.0537	-0.023	0.4211
mkt04y3	0.8475	-0.0213	0.0187	0.281
mkt06y3	-0.0433	0.8486	0.0569	0.2748
mkt08y3	-0.0124	0.0215	0.79	0.3753
mkt09y3	0.2835	-0.1217	0.5051	0.6497
mkt10y3	0.0024	0.0432	0.7472	0.4399
mkt11y3	0.1028	0.6457	0.0595	0.569
mkt12y3	0.2994	0.3518	0.3963	0.6295
mkt14y3	0.4773	0.57	-0.2011	0.4069
mkt15y3	0.6212	0.447	0.0629	0.4103
mkt16y3	0.6313	0.2881	0.2638	0.4489

Figure 5. Year 3 testlet factor loading results, 3-factor solution. Colored variable names indicate testlet groupings

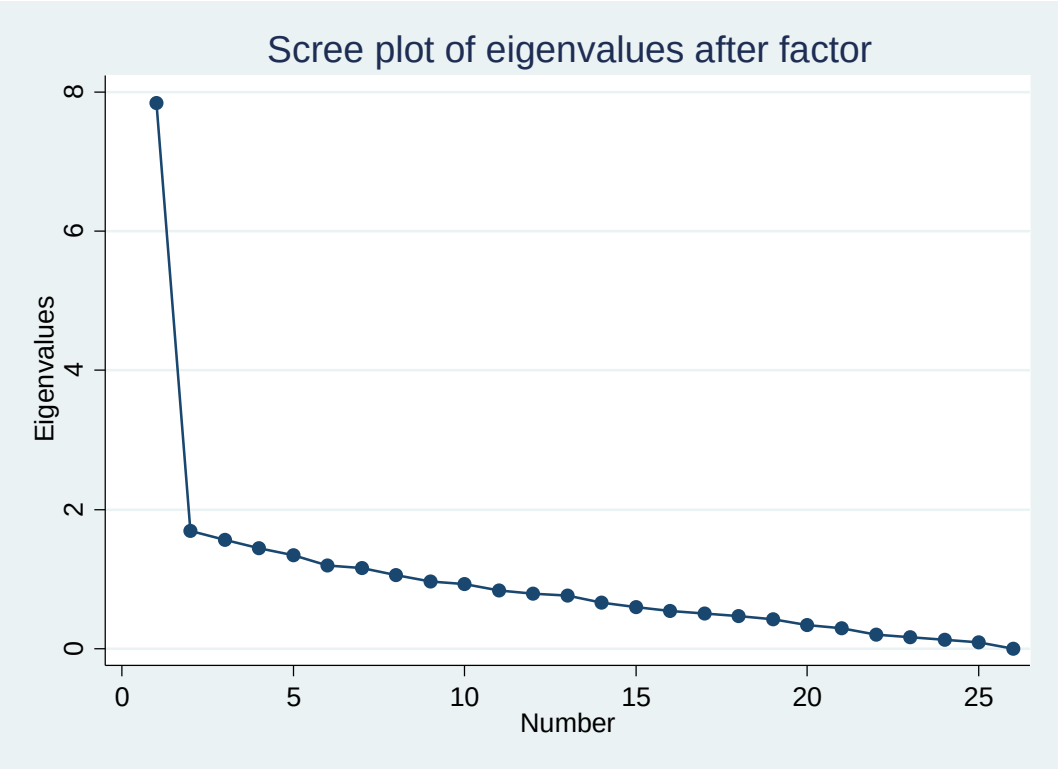


Figure 6. Year 1 EFA scree plot results.

Variable	Factor1	Factor2	Uniqueness
mkt01y1t_4	0.4524	0.269	0.7229
mkt05y1	0.7856	0.1796	0.3506
mkt06y1	0.4844	0.1266	0.7493
mkt07y1	0.6112	0.1725	0.5967
mkt08y1	0.7113	0.1003	0.484
mkt09y1	0.6306	0.2835	0.522
mkt10y1	0.5944	-0.0084	0.6466
mkt11y1	0.4031	0.3086	0.7423
mkt13y1	0.4706	0.1796	0.7463
mkt14y1	0.3564	0.3832	0.7261
mkt15y1	0.5578	0.4209	0.5117
mkt16y1	0.5744	0.2873	0.5876
mkt17y1t2_3	0.5668	0.1626	0.6523
mkt21y1	0.4058	0.4277	0.6525
mkt22y1	0.3629	0.3741	0.7283
mkt23y1t_5	0.4178	0.5058	0.5697
mkt28y1t_4	0.5812	0.2581	0.5955
mkt32y1	0.574	0.1493	0.6483
mtel01y1	0.1245	0.6258	0.5929
mtel02y1	0.2675	0.405	0.7644
mtel03y1	0.3165	0.4518	0.6957
mtel04y1	0.2607	0.3073	0.8376
mtel05y1	0.4734	-0.1864	0.7411
mtel06y1	0.4567	0.1729	0.7616
mtel07y1	0.0322	0.7726	0.4021
mtel08y1	0.1646	0.7346	0.4333

Figure 7. Year 1 EFA factor loading results, 2-factor solution.

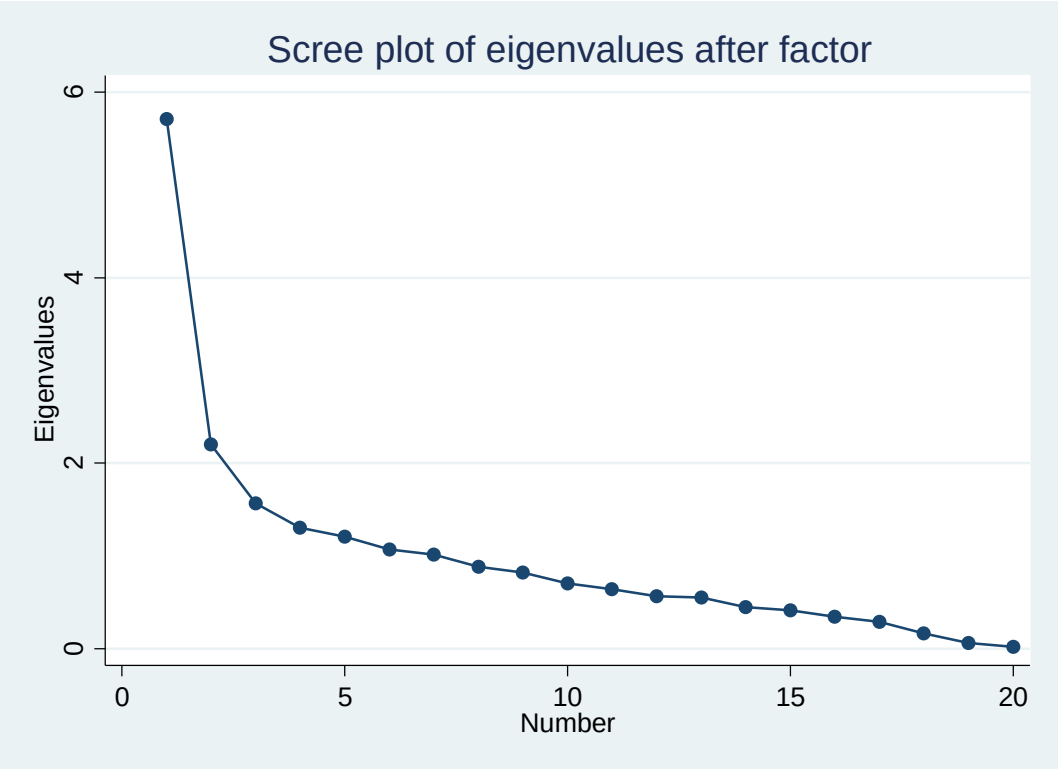


Figure 8. Year 2 EFA scree plot results.

Variable	Factor1	Factor2	Uniqueness
mkt01y2	0.4019	-0.6081	0.4686
mkt02y2	0.388	-0.1231	0.8343
mkt03y2t_3	0.4546	0.1236	0.778
mkt06y2t2_3	0.5838	-0.3159	0.5594
mkt10y2	0.5616	0.2368	0.6286
mkt11y2	0.6397	0.1278	0.5744
mkt12y2	0.2861	0.5157	0.6522
mkt13y2t_5	0.7463	0.0806	0.4366
mkt19y2t2_2	0.4708	0.2778	0.7012
mkt22y2	0.7171	0.0836	0.4788
mkt24y2	0.5227	0.0211	0.7263
mtel02y2	0.7261	-0.1384	0.4536
mtel03y2	0.5952	-0.1515	0.6228
mtel04y2	0.2464	0.5155	0.6736
mtel05y2	0.4517	0.2082	0.7527
mtel06y2	0.5648	0.184	0.6472
mtel08y2	0.0442	0.7413	0.4485
mtel09y2	0.6797	0.0957	0.5289
mtel10y2	0.6167	0.3594	0.4906
mtel11y2	0.1171	0.6005	0.6257

Figure 9. Year 2 EFA factor loading results, 2-factor solution.

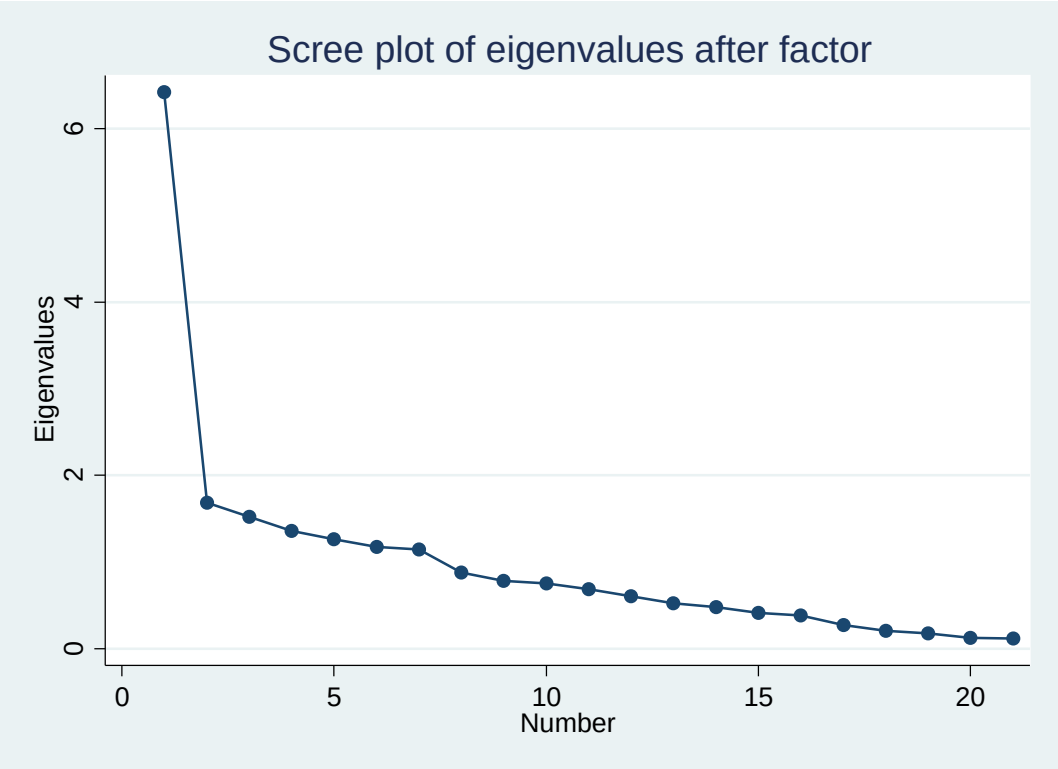


Figure 10. Year 3 EFA scree plot results.

Variable	Factor1	Factor2	Uniqueness
mkt01y3	0.629	0.1902	0.5682
mkt02y3t2_2	0.4074	0.2998	0.7442
mkt06y3	0.1061	0.5147	0.7238
mkt08y3t2_3	0.3811	0.042	0.853
mkt11y3t2_2	0.3121	0.4898	0.6627
mkt14y3	0.1535	0.7797	0.3686
mkt15y3	0.5942	0.3428	0.5294
mkt16y3	0.463	0.4694	0.5653
mtel01y3	0.2789	0.5145	0.6575
mtel02y3	0.0388	0.5722	0.6711
mtel03y3	0.6597	0.1463	0.5433
mtel04y3	0.7544	0.1684	0.4025
mtel05y3	0.096	0.4429	0.7946
mtel06y3	0.4509	0.2396	0.7393
mtel07y3	0.5845	0.2425	0.5995
mtel08y3	0.5991	0.4201	0.4646
mtel09y3	0.1968	0.6911	0.4837
mtel10y3	0.554	0.3465	0.573
mtel11y3	0.6889	0.0705	0.5205
mtel12y3	0.5716	-0.1979	0.6341
mtel13y3	0.2621	0.3721	0.7928

Figure 11. Year 3 EFA factor loading results, 2-factor solution.

	total	mkt	mtel
common	16	1	15
mixed	6	2	4
special	25	24	1

Figure 12. Results from blind categorization of MKT/MTEL items. Common refers to items with a specializedness score of .125 or below. Special refers to items with a specializedness score of .875 or above.

Section 1.a.iii. Fall Questionnaire – Other Teacher Constructs

Math Instruction Richness/Traditional (Stem Code: MIR)

These items were intended to form two separate scales, one representing "traditional math instruction" and the other representing "richness math instruction".

Year 1 analysis recommended two scales:

- Traditional: Y1MIR01, Y1MIR04, Y1MIR08, Y1MIR09.
- Richness: Y1MIR02, Y1MIR03, Y1MIR05, Y1MIR06, Y1MIR07.

Year 1 recommended changes:

- Traditional: replace Y1MIR01 with Y2MIR10; add Y2MIR11 and Y2MIR12.
- Richness: did not recommend any item changes; did recommend combining with Math Instruction Richness items with Math Instruction SPMMR items (below) for an "All Richness" scale.

We followed these recommendations, and additionally dropped Y1MIR07 from the Year 2 TQ.

Year 2 recommended two scales:

- Traditional: Y2MIR10, Y2MIR11, Y2MIR04, Y2MIR08, Y2MIR09, Y2MIR12.
- Richness: Y2MIR02, Y2MIR03, Y2MIR05, Y2MIR06.

Year 2 recommended changes:

- Traditional: Y2MIR10 and Y2MIR04 were problematic: Y2MIR10 cross-loaded with mediocre loadings on both factors; Y1MIR04 loaded only weakly and had low communality (i.e. was a poor fit).
- Richness: no changes.

Based on these recommendations, for Year 3 we removed Y2MIR04 but retained Y2MIR10; we did not add any new items.

Year 3: Year 3 found weak 3-factor structure (third eigenvalue = 1.1); when 2-factor structure is forced, items fall into the expected Traditional (Y2MIR10, Y2MIR11, Y2MIR08, Y2MIR09, Y2MIR12) and Richness (Y2MIR02, Y2MIR03, Y2MIR05, Y2MIR06) factors. Communalities were not great for Y2MIR10 and Y2MIR03, but loadings were fine.

Comparison: Year 1 and Year 2 find the same basic structure, but from year to year we have removed/added items in an effort to improve the scales. For strict comparison between Year 1 and Year 2/Year 3, use:

- Traditional: Only Y1MIR04, Y1MIR08, Y1MIR09 are compatible across both years; but the point of adding items in Year 2 was to increase the number of items in this scale.
- Richness: Y1MIR02, Y1MIR03, Y1MIR05, Y1MIR06.

Math Instruction SPMMR (MIS)

These items were intended to form a single scale. Although we originally intended them to be a separate scale from the MIR items, we also experimented with combining the Richness MIR items with the MIS items for a "total reform-based instruction" scale.

Year 1 recommended a single scale:
Math Instruction SPMMR: Y1MIS02, Y1MIS03, Y1MIS04, Y1MIS05.

Year 1 of this set of items combined with the MIR Richness items recommended a single scale:
Math Instruction Total Reform: Y1MIR02, Y1MIR03, Y1MIR05, Y1MIR06, Y1MIR07, Y1MIS02, Y1MIS03, Y1MIS04, Y1MIS05 (note that item Y1MIR07 was dropped in Y2).

Year 1 recommended changes:
Remove Y1MIS01 and replace it with Y2MIS06.

We followed this recommendation.

Year 2 recommended a single scale:
Math Instruction SPMMR: Y2MIS02, Y2MIS03, Y2MIS04, Y2MIS05, Y2MIS06.

Year 2 recommended changes:
Item Y2MIS03's communality was not great, though it loaded on the factor at > 0.5; removing this item from the scale would increase reliability only a little. We decided to leave this item in the Year 3 survey.

Year 3: Found the expected single-factor solution. Communality of Y2MIS06 was not great, but loading was fine.

Comparison:
Both years find the same structure for Math Instruction SPMMR, but one item was changed/added for Year 2/Year 3. To get the same scale for all three years, use:
Math Instruction SPMMR: Y1MIS02, Y1MIS03, Y1MIS04, Y1MIS05.

Classroom Climate (CLC)

These items were intended to form a unidimensional scale, but half have positive valence and half have negative valence.

Year 1 recommended two scales:

- Positive Classroom Climate: Y1CLC01, Y1CLC05, Y1CLC07, Y1CLC08.
- Negative Classroom Climate: Y1CLC02, Y1CLC04, Y1CLC06, Y1CLC09.

Year 1 recommended changes:
Remove item Y1CLC03.

We followed this recommendation.

Year 2 recommended a single scale:
Y2CLC01, Y2CLC02, Y2CLC04, Y2CLC05, Y2CLC06, Y2CLC07, Y2CLC08, Y2CLC09.

Year 2 recommended no changes.

Year 3 analysis:
Clear 1-factor solution, with all items loading well and only Y3CLC05 having poor communality and causing slight increase to alpha if removed from scale.

Comparison:

- For compatibility across the three years, use: Y1CLC01, Y1CLC02, Y1CLC04, Y1CLC05, Y1CLC06, Y1CLC07, Y1CLC08, Y1CLC09.
- Year 1 recommended two separate scales (positive and negative), Year 2 recommended one.

Formative Assessment (FAS)

These items were intended to form a single unidimensional scale.

Year 1 recommended a single scale:
Formative Assessment: Y1FAS01, Y1FAS02, Y1FAS03, Y1FAS04, Y1FAS05, Y1FAS06
(Note: analysis found a 2-factor structure, with the second factor consisting of Y1FAS04 and Y1FAS05, eigenvalue 1.16; but recommendation was to stick to a single scale).

Year 1 analysis recommended changes:
We removed Y1FAS04 and added Y2FAS07, Y2FAS08, Y1FAS09 for Y2.

Year 2 analysis found 3-factor solution, recommendation not clear:

- Formative Assessment Scoring Rubrics: Y2FAS01, Y2FAS06 (Y2FAS08 cross-loads).
- Formative Assessment Adjusting Instruction: Y2FAS03, Y2FAS07.
- Formative Assessment Understanding Student Thinking: Y2FAS02, Y2FAS05, Y2FAS08, Y2FAS09.

1-factor solution was mediocre in terms of factor loadings and reliability; communalities were poor. 2- or 3-factor solution fit the data better and made conceptual sense.

Year 2 recommended changes:
The initial factor analysis does not speak to recommended changes. What we did was remove this scale from Y3; we retained Y2FAS03, but put it into a new set of items about individualizing instruction (see below).

Comparison:

- Year 1 found weak 2-factor solution and recommended a single scale (reliability 0.69); Year 2 (after addition of new items) found weak 3-factor solution but 1-factor solution did not fit the data well (reliability 0.64).
- For compatibility, use: Y1FAS01, Y1FAS02, Y1FAS03, Y1FAS05, Y1FAS06.
- Comparable scale does not exist for Year 3.

Individualized Instruction (INI)

Year 3: This scale was created for Year 3. It was intended as a single scale and a one-factor solution was found; communality was not great for Y3FAS03, the item recycled from the old Formative Assessment scale, but its loading was fine.

Putting in Time Outside Class (TOC)

Originally intended to be a unidimensional scale.

Year 1 recommended 2 scales:

- Putting In Time Outside Class Math Prep: Y1TOC01, Y1TOC02, Y1TOC03, Y1TOC04.
- Putting In Time Outside Class Other Effort: Y1TOC05, Y1TOC06, Y1TOC07, Y1TOC08.

Year 1 recommended changes:
Drop Y1TOC05, Y1TOC07, Y1TOC08, revise Y1TOC06 to be math-specific.

We followed these recommendations.

Year 2 recommended 1 scale:

- Note that 1 scale is consistent with Year 1 analysis, since we dropped the items that formed the second scale in Y1.
- Putting In Time Outside Class Math Prep: Y2TOC01, Y2TOC02, Y2TOC03, Y2TOC04, Y2TOC06 (revised).

Year 2 recommended changes:
Possibly omit Y2TOC06 (revised) due to mediocre loading and slight increase in reliability. This poorer fit is consistent with Year 1 analysis, but we added the item for a reason.

Comparison:

- Year 1 suggested 2 scales; Year 2 contained the items from 1 of the scales, which showed unidimensionality, with the item adapted from the other scale fitting less well. These analyses were consistent.
- For compatibility, use: Y2TOC01, Y2TOC02, Y2TOC03, Y2TOC04.
- Comparable scale does not exist for Year 3.

Math Leadership Roles (MLR)

Year 1 recommends not treating these items as a scale at all.

Year 1 recommends changes:
Remove item Y1MLR02 due to low variance.

We followed this recommendation.

Year 2 recommends not treating these items as a scale at all.

Year 2 recommends changes:
No changes recommended; these items were not used in Year 3 TQ.

Test Prep Activities (TPA)

Year 1:
We decided to retain the variables, but re-wrote them.

Year 2 recommended a single scale:
Test Prep Activities: Y2TPA01, Y2TPA02, Y2TPA03, Y2TPA04, Y2TPA05.

Year 2 recommended no changes.

Test Prep Change (TPC)

Year 1:
We decided to retain the variables, but re-write them.

Year 2 recommended a single scale:
Test Prep Activities: Y2TPC01, Y2TPC02, Y2TPC03, Y2TPC04, Y2TPC05, Y2TPC06, Y2TPC07 (weak second factor found but properties of 1-factor solution good).

Year 2 recommended no changes.

Year 3 Test Prep (TPA/TPC)

Year 3:
Test prep items were heavily revised between Year 2 and Year 3, so this analysis treated the Year 3 items as a new scale, incompatible with previous years. The items were collected under a single stem for Year 3. Solid 2-factor solution, with the breakdown almost, but not quite according, to the original two stems: TPC02 and TPC07 load with the TPAs, and there's a bit of cross-loading. This makes some conceptual sense: the other TPCs are about "spending less time" and these two are about focusing instruction on stuff that's on the test, which is similar to the TPA questions.

Students This Year (Compared to Last) (STY)

These items were intended as a unidimensional scale.

Year 1 recommended a single scale.

Year 1 recommended no changes.

Year 2 recommended a single scale.

Year 2 recommended no changes.

Year 3 analysis: found expected 1-factor structure; Y3STY01 raises alpha by a tiny amount if removed from scale, but uniqueness and loading are fine.

Comparison: Compatible

Teacher External Blame / Teacher Self-Efficacy (TEB/TSE)

These items were initially intended as a unidimensional scale, but were pulled from two sources/constructs and the factor structure reflected that fact.

Year 1 recommended 2 scales:

- Teacher Self-Efficacy: Y1TSE02, Y1TSE03, Y1TSE04, Y1TSE05.
- External Blame: Y1TEB01, Y1TEB02, Y1TEB03.

Year 1 analysis recommended changes:

- Teacher Self-Efficacy: Remove item Y1TSE01, fix response options to unweight from 'agree'.
- External Blame: build longer/better scale.

Based on these recommendations, we changed the number of response options, stem, and item wording for all the items, so that corresponding items from Year 1 and Year 2 are no longer equivalent. We removed item Y1TSE01, as recommended, and also item Y1TSE04; and we added a large number of new Self-Efficacy items. We removed most of the External Blame items, replacing them with Y2TEB04, Y2TEB05, Y2TEB06, Y2TEB07. We also split the External Blame and Self-Efficacy items into two separate stems on the survey.

Year 2 analysis recommended:
In Year 2, the External Blame and Self-Efficacy items were under two separate stems and analyzed separately.

- Self-Efficacy, 2 scales:
    - Self-Efficacy: Effect On Students: Y2TSE06, Y2TSE07, Y2TSE08, Y2TSE09, Y2TSE010, Y2TSE011, Y2TSE12.
    - Self-Efficacy: Teaching Strategies: Y2TSE02, Y2TSE03, Y2TSE05, Y2TSE13, Y2TSE14.
- External Blame, 1 scale: Y2TEB04, Y2TEB05, Y2TEB06, Y2TEB07.

Year 2 analysis recommended changes:

- No particular changes recommended.
- These items were not included in Year 3.

Comparison:

- There's really no way to compare these scales across years; too much editing was done for comparable scales to be created.
- These items were not used in Year 3.

School Provides Resources (SPR)

Items originally intended as a single scale.

Year 1 recommended a single scale, but 2 alternative versions:

- School Provides Resources: Y1SPR01, Y1SPR02, Y1SPR03, Y1SPR05, Y1SPR06, Y1SPR07, Y1SPR08, Y1SPR09.
- Teacher Likes The School: Y1SPR03, Y1SPR04, Y1SPR05, Y1SPR06, Y1SPR07, Y1SPR08, Y1SPR09.

Year 1 recommended no changes.

Year 2:
Found weak 3-factor solution; there are arguments for either 2 scales or 1; we didn't make a decision.

- Materials/resources: Y2SPR01, Y2SPR05, Y2SPR06, Y2SPR07, Y2SPR08 (note cross-loads).
- Emotional/environment: Y2SPR02, Y2SPR03, Y2SPR07, Y2SPR09.

Year 2 recommended:

- In the 1 and 2 factor solutions, Y2SPR04 does not load well.
- However, we made no changes in Year 3.

Year 3:
We ran analysis including item #4. Weak 3-factor solution (with cross-loading):

- Y3SPR01, Y3SPR05, Y3SPR06, (Y3SPR07, Y3SPR08).
- Y3SPR02, Y3SPR03, (Y3SPR06), Y3SPR09.
- Y3SPR04, Y3SPR07, Y3SPR08.

When 2-factor solution forced:

- Y3SPR01, Y3SPR05, Y3SPR06, (Y3SPR07, Y3SPR08).
- Y3SPR02, Y3SPR03, Y3SPR04, (Y3SPR07, Y3SPR08), Y3SPR09.

Similar, though not completely identical, to Year 2.

Comparison:

- Item Y2SPR04 was ill-fitting in both years, so one may want to exclude it from analysis, though we included it on the survey in Year 3.
- Single scale: Y1SPR01, Y1SPR02, Y1SPR03, Y1SPR05, Y1SPR06, Y1SPR07, Y1SPR08, Y1SPR09 (if we exclude Y2SPR04).

Support from Coaches (SFC)

Year 2 recommends 1 scale:
Found 2 factors, but the 1-factor solution made conceptual sense, gave decent reliability, and demonstrated good loadings except for Y2SFC01.

Year 2 recommends no changes:
We kept these items unchanged for Year 3.

Year 3:
Weak 2- factor solution:

- Y3SFC01, Y3SFC02, Y3SFC06.
- Y3SFC03, Y3SFC04, Y3SFC05.

In a forced 1-factor solution, loadings were good, uniquenesses was not great and was pretty poor for Y3SFC03, Y3SFC04, Y3SFC06. We decided to stick to single scale.

Comparison:
For compatibility, use: Y3SFC01, Y3SFC02, Y3SFC03, Y3SFC04, Y3SFC05.

Teacher Professional Development (TPD)

Year 2:
Found 2 factors that make conceptual sense, but 1-factor solution has reasonable-to-strong loadings and reliability of 0.84.

Year 2 recommends no changes:
We kept these items unchanged for Year 3.

Year 3:
TPD01-TPD07 form a single factor, with 07 not loading very well and demonstrating poor communality. Removing it would raise alpha a little.

Year 3 recommends:
Make a single scale, and also a math (01-04) and non-math (05-07) scale, as in Year 2.

Perceived Usefulness of Evaluator Feedback (POF)

Year 3:
2 factors, but second eigenvalue is 1.3. Loadings were good for 1-factor, though uniquenesses were so-so.

- Feedback encouraged change in practice (POF11, POF14, POF16).
- Feedback was pertinent (POF12, POF13, POF15).

Year 3 recommends:
Create a single scale with all variables, and also the two separate scales

- Perception of Feedback: Y3POF11, Y3POF12, Y3POF13, Y3POF14, Y3POF15, Y3POF16.
- Feedback encouraged change: Y3POF11, Y3POF12, Y3POF13.
- Feedback was pertinent: Y3POF14, Y3POF15, Y3POF16.

School Leadership (STL)

Year 2 (Spring):
2-factors, second eigenvalue 1.2.

- how teachers are perceived/respected (STL01-STL03).
- leadership/community practices (STL04-STL07).

Year 2 recommends:
Single scale recommended.

Year 3 (Fall and Spring):
Both Fall and Spring Year 3 found unproblematic single scale.

Comparison:
Single scale: Y2STL01, Y2STL02, Y2STL03, Y2STL04, Y2STL05, Y2STL06, Y2STL07.

School Community/Collaboration (SCC)

Year 2 (Spring):
Single scale, with communality issues for 01 and 05 but loadings and alphas okay.

Year 3 (Fall and Spring):
In fall found marginal 2-factor, in Spring only 1-factor; in both cases, 01 and 05 had not-great communalities, but loadings and alphas okay.

Year 2/Year 3 recommends:
Single scale: Y2SCC01, Y2SCC02, Y2SCC03, Y2SCC04, Y2SCC05, Y2SCC06, Y2SCC07, Y2SCC08.

Cronbach's Alphas

Year 1

```
 **** variable name:  trad_TQ10 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .2602054
Number of items in the scale:            4
Scale reliability coefficient:      0.5645
 description: traditional instruction


 **** variable name:  rich_TQ10 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .2635869
Number of items in the scale:            5
Scale reliability coefficient:      0.6900
 description: rich instruction


 **** variable name:  trad_compatible_TQ10

Test scale = mean(unstandardized items)

Average interitem covariance:     .3172955
Number of items in the scale:            3
Scale reliability coefficient:      0.5434
 description: traditional instruction, cross-year compatible


 **** variable name:  rich_compatible_TQ10 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .3050431
Number of items in the scale:            4
Scale reliability coefficient:      0.6791
 description: rich instruction, cross-year compatible


 **** variable name:  SPMMR_TQ10 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .4402699
Number of items in the scale:            4
Scale reliability coefficient:      0.7530
 description: SPMMR instruction


 **** variable name:  SPMMR_compatible_TQ10 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .4402699
Number of items in the scale:            4
Scale reliability coefficient:      0.7530
 description: SPMMR instruction, cross-year compatible


 **** variable name:  allreform_TQ10 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .3098704
Number of items in the scale:            9
```

```
Scale reliability coefficient:      0.8186
 description: reform-based mathematics instruction


 **** variable name:  negclim_TQ10 ****

Test scale = mean(unstandardized items)

Average interitem covariance:      .515616
Number of items in the scale:            4
Scale reliability coefficient:      0.8861
 description: negative classroom climate


 **** variable name:  posclim_TQ10 ****

Test scale = mean(unstandardized items)

Average interitem covariance:      .384945
Number of items in the scale:            4
Scale reliability coefficient:      0.7893
 description: positive classroom climate


 **** variable name:  totalclim_TQ10 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .3962807
Number of items in the scale:            8
Scale reliability coefficient:      0.8899
 description: total classroom climate


 **** variable name:  formass_TQ10 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .2297084
Number of items in the scale:            6
Scale reliability coefficient:      0.6906
 description: use of formative assessment


 **** variable name:  formass_compatible_TQ10 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .2747657
Number of items in the scale:            5
Scale reliability coefficient:      0.6770
 description: use of formative assessment, cross-year compatible


 **** variable name:  mathprep_compatible_TQ10 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .3006021
Number of items in the scale:            4
Scale reliability coefficient:      0.7895
 description: mathematics class preparation activities, cross-year compatible


 **** variable name:  otheffort_TQ10 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .2206841
Number of items in the scale:            4
Scale reliability coefficient:      0.5718
```

```
 description: other preparation activities


 **** variable name:  testprepact_TQ10 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .3132512
Number of items in the scale:            5
Scale reliability coefficient:      0.7387
 description: use of test prep activities


 **** variable name:  testprepchg_TQ10 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .6773682
Number of items in the scale:            7
Scale reliability coefficient:      0.8610
 description: changing instruction for test prep


 **** variable name:  studworse_TQ10 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .8420731
Number of items in the scale:            4
Scale reliability coefficient:      0.7617
 description: perceptions of students being relatively worse than last year


 **** variable name:  teacheff_TQ10 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .1459571
Number of items in the scale:            4
Scale reliability coefficient:      0.6587
 description: teacher self-efficacy


 **** variable name:  extblame_TQ10 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .3796038
Number of items in the scale:            3
Scale reliability coefficient:      0.5584
 description: external blame


 **** variable name:  resources_TQ10 ****

Test scale = mean(unstandardized items)

Average interitem covariance:      .456582
Number of items in the scale:            9
Scale reliability coefficient:      0.7994
 description: school provides resources
```

Year 2

```
 **** variable name:  trad_TQ11 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .3325834
Number of items in the scale:            6
Scale reliability coefficient:      0.7254
```

```
 description: traditional instruction


 **** variable name:  rich_TQ11 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .2867677
Number of items in the scale:            4
Scale reliability coefficient:      0.6851
 description: rich instruction


 **** variable name:  trad_compatible_TQ11 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .2506339
Number of items in the scale:            3
Scale reliability coefficient:      0.4802
 description: traditional instruction, cross-year compatible


 **** variable name:  rich_compatible_TQ11 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .2867677
Number of items in the scale:            4
Scale reliability coefficient:      0.6851
 description: rich instruction, cross-year compatible


 **** variable name:  SPMMR_TQ11 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .3722069
Number of items in the scale:            5
Scale reliability coefficient:      0.7751
 description: SPMMR instruction


 **** variable name:  SPMMR_compatible_TQ11 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .3906052
Number of items in the scale:            4
Scale reliability coefficient:      0.7349
 description: SPMMR instruction, cross-year compatible


 **** variable name:  negclim_TQ11 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .3978216
Number of items in the scale:            4
Scale reliability coefficient:      0.8450
 description: negative classroom climate


 **** variable name:  posclim_TQ11 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .3149612
Number of items in the scale:            4
Scale reliability coefficient:      0.7777
 description: positive classroom climate
```

```
 **** variable name:  totalclim_TQ11 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .3294777
Number of items in the scale:            8
Scale reliability coefficient:      0.8819
 description: total classroom climate


 **** variable name:  formass_ass_TQ11 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .2716092
Number of items in the scale:            4
Scale reliability coefficient:      0.6152
 description: use of formative assessment - assessment


 **** variable name:  formass_inst_TQ11 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .1730437
Number of items in the scale:            3
Scale reliability coefficient:      0.5684
 description: use of formative assessment - instruction


 **** variable name:  formass_compatible_TQ11 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .1468554
Number of items in the scale:            5
Scale reliability coefficient:      0.5266
 description: use of formative assessment, cross-year compatible


 **** variable name:  mathprep_TQ11 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .2589076
Number of items in the scale:            5
Scale reliability coefficient:      0.7250
 description: mathematics class preparation activities


 **** variable name:  mathprep_compatible_TQ11 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .2913242
Number of items in the scale:            4
Scale reliability coefficient:      0.7671
 description: mathematics class preparation activities, cross-year compatible


 **** variable name:  testprepact_TQ11 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .3548945
Number of items in the scale:            5
Scale reliability coefficient:      0.7961
 description: use of test prep activities
```

**** variable name: testprepchg_TQ11 ****

Test scale = mean(unstandardized items)

Average interitem covariance: .7262751
Number of items in the scale: 7
Scale reliability coefficient: 0.8692
 description: changing instruction for test prep

**** variable name: studworse_TQ11 ****

Test scale = mean(unstandardized items)

Average interitem covariance: 1.253526
Number of items in the scale: 4
Scale reliability coefficient: 0.8482
 description: perceptions of students being relatively worse than last year

**** variable name: teacheff_students_TQ11 ****

Test scale = mean(unstandardized items)

Average interitem covariance: .5598631
Number of items in the scale: 7
Scale reliability coefficient: 0.8785
 description: teacher self-efficacy on students

**** variable name: teacheff_strategies_TQ11 ****

Test scale = mean(unstandardized items)

Average interitem covariance: .4049935
Number of items in the scale: 5
Scale reliability coefficient: 0.7310
 description: teacher self-efficacy on teaching strategies

**** variable name: extblame_TQ11 ****

Test scale = mean(unstandardized items)

Average interitem covariance: .9163763
Number of items in the scale: 4
Scale reliability coefficient: 0.9278
 description: external blame

**** variable name: resources_TQ11 ****

Test scale = mean(unstandardized items)

Average interitem covariance: .3993205
Number of items in the scale: 9
Scale reliability coefficient: 0.7750
 description: school provides resources

**** variable name: coachsupport_TQ11 ****

Test scale = mean(unstandardized items)

Average interitem covariance: .4795874
Number of items in the scale: 5
Scale reliability coefficient: 0.7492
 description: amount of support from coaches

```
 **** variable name:  coachsupport_compatible_TQ11 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .4795874
Number of items in the scale:            5
Scale reliability coefficient:      0.7492
 description: amount of support from coaches, cross-year compatible


 **** variable name:  pdmath_TQ11 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .7768121
Number of items in the scale:            4
Scale reliability coefficient:      0.8363
 description: amount of math PD in past year


 **** variable name:  pdother_TQ11 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .5068833
Number of items in the scale:            3
Scale reliability coefficient:      0.6685
 description: amount of non-math PD in past year


 **** variable name:  pdgeneral_TQ11 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .6593789
Number of items in the scale:            6
Scale reliability coefficient:      0.8541
 description: amount of PD overall in past year
```

Year 3

```
 **** variable name:  trad_TQ12 ****
MIR04 constant in analysis sample, dropped from analysis

Test scale = mean(unstandardized items)

Average interitem covariance:     .4700339
Number of items in the scale:            5
Scale reliability coefficient:      0.7825
 description: traditional instruction


 **** variable name:  rich_TQ12 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .2811689
Number of items in the scale:            4
Scale reliability coefficient:      0.6897
 description: rich instruction


 **** variable name:  rich_compatible_TQ12 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .2811689
Number of items in the scale:            4
Scale reliability coefficient:      0.6897
 description: rich instruction, cross-year compatible
```

```
 **** variable name:  SPMMR_TQ12 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .4121006
Number of items in the scale:            5
Scale reliability coefficient:      0.8153
 description: SPMMR instruction


 **** variable name:  SPMMR_compatible_TQ12 ****

Test scale = mean(unstandardized items)

Average interitem covariance:      .489671
Number of items in the scale:            4
Scale reliability coefficient:      0.8084
 description: SPMMR instruction, cross-year compatible


 **** variable name:  negclim_TQ12 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .5132082
Number of items in the scale:            4
Scale reliability coefficient:      0.8664
 description: negative classroom climate


 **** variable name:  posclim_TQ12 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .3008799
Number of items in the scale:            4
Scale reliability coefficient:      0.7518
 description: positive classroom climate


 **** variable name:  totalclim_TQ12 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .3800051
Number of items in the scale:            8
Scale reliability coefficient:      0.8878
 description: total classroom climate


 **** variable name:  indinstruc_TQ12 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .3391485
Number of items in the scale:            4
Scale reliability coefficient:      0.6913
 description: individualized instruction


 **** variable name:  testprep_revised_TQ12 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .3147093
Number of items in the scale:           11
Scale reliability coefficient:      0.8227
 description: test prep overall
```

**** variable name: testprepact_revised_TQ12 ****

Test scale = mean(unstandardized items)

Average interitem covariance: .4593364
Number of items in the scale: 5
Scale reliability coefficient: 0.7875
 description: use of test prep activities

**** variable name: testprepchg_revised_TQ12 ****

Test scale = mean(unstandardized items)

Average interitem covariance: .3292616
Number of items in the scale: 6
Scale reliability coefficient: 0.7339
 description: changing instruction for test prep

**** variable name: studworse_TQ12 ****

Test scale = mean(unstandardized items)

Average interitem covariance: 1.098639
Number of items in the scale: 4
Scale reliability coefficient: 0.8429
 description: perceptions of students being relatively worse than last year

**** variable name: resources_TQ12 ****

Test scale = mean(unstandardized items)

Average interitem covariance: .3712309
Number of items in the scale: 9
Scale reliability coefficient: 0.7528
 description: school provides resources

**** variable name: coachsupport_TQ12 ****

Test scale = mean(unstandardized items)

Average interitem covariance: .4515816
Number of items in the scale: 5
Scale reliability coefficient: 0.7461
 description: amount of support from coaches

**** variable name: coachsupport_compatible_TQ12 ****

Test scale = mean(unstandardized items)

Average interitem covariance: .4515816
Number of items in the scale: 5
Scale reliability coefficient: 0.7461
 description: amount of support from coaches, cross-year compatible

**** variable name: pdmath_TQ12 ****

Test scale = mean(unstandardized items)

Average interitem covariance: .7924828
Number of items in the scale: 4
Scale reliability coefficient: 0.8435
 description: amount of math PD in past year

```
 **** variable name:  pdother_TQ12 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .3713257
Number of items in the scale:            3
Scale reliability coefficient:      0.5346
 description: amount of non-math PD in past year


 **** variable name:  pdgeneral_TQ12 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .6690939
Number of items in the scale:            6
Scale reliability coefficient:      0.8555
 description: amount of PD overall in past year


 **** variable name:  evalfeedqual_TQ12 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .5749906
Number of items in the scale:            6
Scale reliability coefficient:      0.7974
 description: perceived usefulness of evaluator feedback


 **** variable name:  evalfeedchg_TQ12 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     1.073057
Number of items in the scale:            3
Scale reliability coefficient:      0.8140
 description: evaluator feedback encouraged change


 **** variable name:  evalfeedpert_TQ12 ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .5793249
Number of items in the scale:            3
Scale reliability coefficient:      0.7715
 description: evaluator feedback was pertinent


 **** variable name:  Y3F_STL_Scale ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .6953007
Number of items in the scale:            7
Scale reliability coefficient:      0.9200
 description: school teacher leadership


 **** variable name:  Y3F_SCC_Scale ****

Test scale = mean(unstandardized items)

Average interitem covariance:     .4002795
Number of items in the scale:            8
Scale reliability coefficient:      0.8534
 description: school community/collaboration
```

Section 1.b.i. Spring Questionnaire – Overview

NCTE administered surveys to participating teachers in the spring of each data-collection year. In addition to items assessing teachers' *Knowledge of Students* (KOS), the spring questionnaires contained a few sets of items about teachers' coverage of mathematics content and perceptions of their school environments. The items on the spring questionnaire relating to school environments appeared only on surveys in years 2 and 3.

For the questionnaire, exploratory factor analysis (EFA) of the teachers' responses to the school environment items was used to determine the factor structure of each set of survey items designed to measure a latent construct. The results of the EFA informed the creation of composite scales to be used in analysis, as well as decisions about revising, deleting, or adding items to the following year's survey.

In year one of NCTE, 252 of 263 spring surveys sent out were completed, for a response rate of 96%. In year two, this rate was 98% (214/219), and in year three, this rate was 100% (115/115). In what follows, we describe the technical aspects of the construction of various constructs captured in the questionnaire.

Section 1.b.ii. Spring Questionnaire – *Knowledge of Students* (KOS)

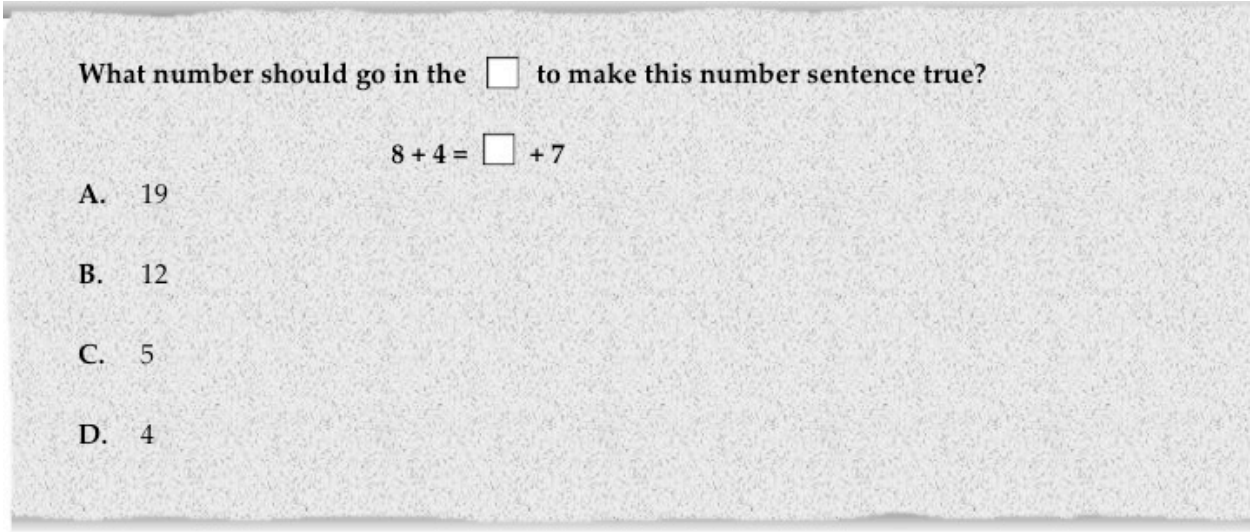
Note: The following passage is excerpted from the following manuscript:

Hill, H. C., & Chin, M. (submitted). *Teacher's knowledge of students: Defining a domain*.

Methods

As noted above, the spring teacher questionnaires contained questions intended to assess two different aspects of teachers' knowledge of students (KOS). One, knowledge of student misconceptions (KOSM), reflects a component contained within Shulman's (1986, 1987) pedagogical content knowledge and Ball and colleagues' (2008) KCS categories. To measure KOSM, we followed the strategy used by Sadler and colleagues (2013). Specifically, the questionnaire reprinted items from the project-developed mathematics test and asked teachers, "Which [of the following options for this item] will be the most common incorrect answer among fourth [or fifth] graders **in general**?"[1] Project staff selected the student test items to place on the questionnaire strategically, attempting to exclude those for which there was not a dominant incorrect student response and prioritizing items for which the dominant student response was well established in the research literature.

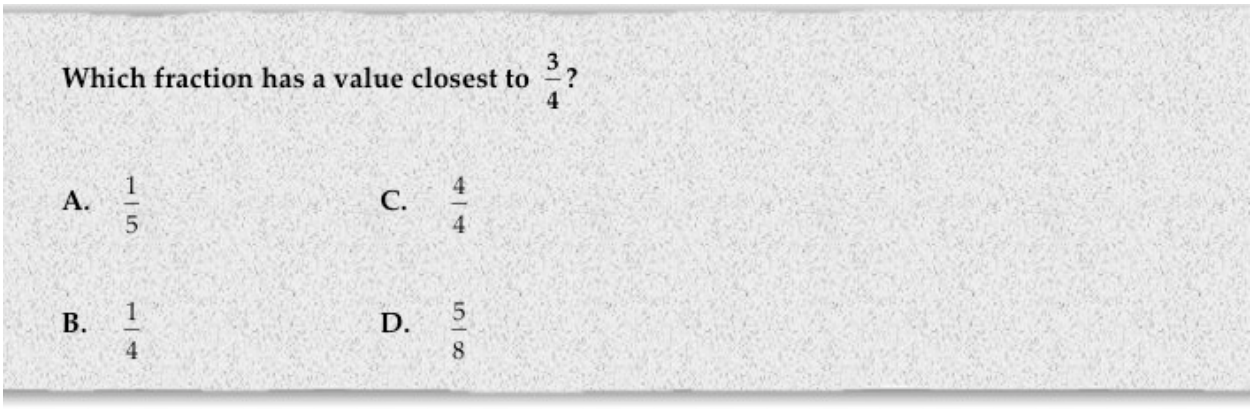
Consider the following problem from the student assessment:



The correct answer to this problem is C. a. Approximately what percentage of **your students** being tested today will choose the correct answer?	%
b. Approximately what percentage of fourth grade students in **your district** will choose the correct answer?	%
c. Which will be the most common incorrect answer among fourth graders **in general**? (Please circle **ONE** answer.)	A B D

Figure 1. Example item on spring teacher questionnaire used to assess both accuracy and KOSM. For the KOSM measure, this item has a researched-aligned dominant incorrect student response.

Consider the following problem from the student assessment:



The correct answer to this problem is D. a. Approximately what percentage of **your students** being tested today will choose the correct answer?	%
b. Approximately what percentage of fifth grade students in **your district** will choose the correct answer?	%
c. Which will be the most common incorrect answer among fifth graders **in general**? (Please circle **ONE** answer.)	A B C

Figure 2. Example item on spring teacher questionnaire used to assess both accuracy and KOSM. For the KOSM measure, this item has a simple dominant incorrect student response.

For instance, the first sample item, depicted in Figure 1, probed students' understanding of the equal sign—as an indicator to compute (yielding the most common incorrect answer of 12) or as an indicator that the quantities on either side of the sign are equal to one another (yielding the correct answer of 5). This student misconception is well documented in the research literature (Knuth, Stephens, MacNeil, & Alibali, 2006). Because the project-administered test did not contain enough items reflecting common student misconceptions,[2] however, many student test items simply had a dominant wrong answer, as in Figure 2, where 69% of answering fifth graders incorrectly chose "C". Total, the spring questionnaires included 21 fourth-grade and 20 fifth-grade KOSM items distributed roughly equally across two study years (2010–11; 2011–12).

To generate KOSM scores for our reliability analysis, we first compared teachers' responses to each question to the actual modal incorrect response of fourth or fifth graders.[3] We then estimated the following one-parameter logistic IRT model within grade (as knowledge of student items differed across grades) using the gsem command in Stata (version 13.1):

$$P(y_{it} = 1|\theta_t, \alpha_i) = logistic(m\theta_t - \alpha_i) \qquad (1)$$

In Equation 1, y_{it} indicates whether teacher t correctly predicted the modal incorrect response among students for item i of the project-developed mathematics test, controlling for α_i, the item difficulty. From Equation 1, we recovered each teacher's "career" (multi-year) KOSM score, θ_t, generated using his or her responses to all KOSM items. We also estimated Equation 1 within school year and grade to recover within-year KOSM scores to use in several analyses below.

We modeled the second measure of teachers' knowledge of students on educational psychologists' notions of judgment *accuracy*, or the extent to which teachers can predict student performance on specified material. To measure this construct, we used the same student items as for KOSM as well as new items in the third study year (2012–13), this time asking teachers, "Approximately what percentage of **your students** being tested today will choose the correct answer [for this item]?"[4] Both fourth- and fifth-grade teachers answered 37 such items total, with items distributed roughly equally across the three years of the study.

To generate accuracy scores for our reliability analyses, we calculated the actual percentage of correct student answers for each item, addressed potential ceiling and floor effects by transforming both the predicted and actual percentages into logits, and then differenced the two values.[5] We then used the absolute values of these differences in the following multilevel equation:

$$y_{it} = \beta_0 + \alpha_i + \theta_t + \varepsilon_{it} \qquad (2)$$

The outcome in Equation 2 represents this absolute difference between predicted and actual logits on item *i* for teacher t. The model also includes a vector of item fixed effects, α_i, capturing differences in item difficulty to correct for the mix of items taken by a specific teacher, and teacher random effects, θ_t representing teachers' underlying accuracy scores.[6] In addition to estimating accuracy scores within grade from items across all years of the study ("career" scores), we also estimated Equation 2 within school year to recover within-year scores for analyses. We multiplied all scores for the accuracy measure by -1 so that higher scores reflected more accurate predictions.

Differentiating among teachers. To ascertain the extent to which these metrics differentiated among teachers, we estimated three reliability metrics. For the KOSM measure, we used estimates of the marginal reliability produced in the IRT model described above. The marginal reliability statistic compares the variance of teacher scores to the expected value of error variance of scores and is comparable to ICCs of classical test theory (see Sireci, Thissen, & Wainer, 1991). For the accuracy measure, we estimated the signal-to-noise ratio in teacher scores using the ICC statistic, adjusted for the average number of items answered by teachers within each grade. Finally, we also used the within-year estimates of accuracy and KOSM scores to examine cross-year correlations, a measure of consistency.

When conducting these analyses, we used the largest sample of teachers possible (i.e., the sample of all teachers who responded to the knowledge of student items on any spring questionnaire): 315 and 306 teachers for accuracy and KOSM, respectively. Results are presented separately by grade, as the student items used were specific to each grade level.

Results

Using adjusted ICCs, we estimated the reliability of fourth-grade accuracy scores to be 0.74 and fifth-grade scores to be 0.72. We also investigated the adjusted ICCs for the set of 22 fourth-grade teachers and 25 fifth-grade teachers who responded to all items ($N = 37$) measuring accuracy. The adjusted ICCs for these samples were 0.76 and 0.77, respectively. These values suggest that, with enough item responses, our measure of accuracy can reliably differentiate among teachers' performance on this construct.

The marginal reliability statistic produced following the estimation of the KOSM IRT model was 0.21 for fourth-grade teachers and 0.40 for fifth-grade teachers. The magnitude of the average standard errors of scores reflected these reliability coefficients, suggesting high imprecision for the average individual; for fourth-grade teachers, the average magnitude of the standard error of KOSM scores was 0.93 *SD*, and for fifth grade, the average magnitude was 0.85. The low estimates of reliability and the noisiness of score estimates suggested that the KOSM measure did not adequately differentiate teachers.

Constructing and then correlating within-year scores provided additional evidence on the reliability of scores. Depending on the grade and combination of years, the cross-year correlation of accuracy scores was moderate, ranging from 0.29 to 0.53; these estimates suggested that teachers' ability to predict the proficiency of their students was somewhat consistent from school year to school year, despite changes in the students taught. Furthermore, correlations were in the same range as the cross-year correlations of other measures of teacher quality (Goldhaber & Hansen, 2013; McCaffrey, Sass, Lockwood, & Mihaly, 2009; Polikoff, 2015). KOSM scores, as expected given their estimated overall reliability, demonstrated less consistency from one year to the next. For fourth-grade teachers, scores correlated at 0.22 between 2010–11 and 2011–12; for fifth-grade teachers, this correlation was slightly higher, at 0.26.

References

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, *59*(5), 389–407.

Goldhaber, D., & Hansen, M. (2013). Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *Economica*, *80*(319), 589–612.

Knuth, E. J., Stephens, A. C., McNeil, N. M., & Alibali, M. W. (2006). Does understanding the equal sign matter? Evidence from solving equations. *Journal for Research in Mathematics Education, 37*(4), 297–312.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*(4), 572–606.

Sadler, P. M., Sonnert, G., Coyle, H. P., Cook-Smith, N., & Miller, J. L. (2013). The influence of teachers' knowledge on student learning in middle school physical science classrooms. *American Educational Research Journal, 50*(5), 1020–1049.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational researcher, 15*(2), 4–14.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard educational review, 57*(1), 1–23.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the Reliability of Testlet-Based Tests. *Journal of Educational Measurement, 28*(3), 237–247.

Section 1.b.iii. Spring Questionnaire – Other Teacher Constructs

School Leadership (STL)

Year 2 (Spring):
2-factors, second eigenvalue 1.2.

- how teachers are perceived/respected (STL01-STL03).
- leadership/community practices (STL04-STL07).

Year 2 recommends:
Single scale recommended.

Year 3 (Fall and Spring):
Both Fall and Spring Year 3 found unproblematic single scale.

Comparison:
Single scale: Y2STL01, Y2STL02, Y2STL03, Y2STL04, Y2STL05, Y2STL06, Y2STL07.

School Community/Collaboration (SCC)

Year 2 (Spring):
Single scale, with communality issues for 01 and 05 but loadings and alphas were okay.

Year 3 (Fall and Spring):
In Fall found marginal 2-factor, in Spring only 1-factor; in both cases, 01 and 05 had not-great communalities, but loadings and alphas were okay.

Year 2/Year 3 recommends:
Single scale: Y2SCC01, Y2SCC02, Y2SCC03, Y2SCC04, Y2SCC05, Y2SCC06, Y2SCC07, Y2SCC08.

Cronbach's Alphas

Year 2

```
 ***** variable: Y2SP_STL_Scale *****

Test scale = mean(unstandardized items)

Average interitem covariance:     13889.13
Number of items in the scale:            7
Scale reliability coefficient:      0.9822
 description: school teacher leadership


 ***** variable: Y2SP_SCC_Scale *****

Test scale = mean(unstandardized items)

Average interitem covariance:     4597.231
Number of items in the scale:            8
Scale reliability coefficient:      0.9409
 description: school community/collaboration
```

Year 3

```
 ***** variable: Y3SP_STL_Scale *****

Test scale = mean(unstandardized items)

Average interitem covariance:     .7018343
Number of items in the scale:            7
Scale reliability coefficient:      0.9337
 description: school teacher leadership


 ***** variable: egen Y3SP_SCC_Scale *****

Test scale = mean(unstandardized items)

Average interitem covariance:     8534.339
Number of items in the scale:            8
Scale reliability coefficient:      0.9547
 description: school community/collaboration
```

Section 2. Observation Instruments - *Mathematical Quality of Instruction* (*MQI*) and *Classroom Assessment Scoring System* (*CLASS*)

Section 2.a. Observation Instruments – Overview

Teachers participating in NCTE were scheduled to be video-recorded three times a year during instruction of a mathematics lesson. Teachers were able to choose when to record themselves, with one exception: they were asked to not record lessons that heavily involved preparing students for tests or that involved students actually taking tests. Lessons typically lasted about an hour in length.

These video-recorded lessons were then scored on the MQI observation instrument—a mathematics-specific observation rubric (the instrument itself can be found in other parts of the documentation for this data deposit). Each lesson was broken down into 7.5-minute segments for scoring, and was scored on the entire instrument by two trained external raters. These raters passed certification exams to score the MQI, and attended biweekly calibration meetings to ensure standardization of scoring procedures. Assignment of raters to lessons minimized the number of instances where a rater scored the same teacher.

These video-recorded lessons were also scored on the CLASS observation instrument—a subject-independent observation rubric (samples from the instrument itself can be found in other parts of the documentation for this data deposit). Each lesson was broken down into 15-minute segments for scoring, and was scored on the entire instrument by a single trained external rater. Raters attended biweekly calibration meetings to ensure standardization of scoring procedures. Assignment of raters to lessons minimized the number of instances where a rater scored the same teacher. For more information on the CLASS, please see http://curry.virginia.edu/research/centers/castl/class.

Not every teacher in the study participated in all three years, and, in some instances, due to poor video capture, lessons were unusable. Thus, not every teacher received MQI or CLASS scores from assigned raters on nine lessons. In this section, we describe select psychometric properties of the two instruments.

Section 2.b. Observation Instruments – Factor Analysis

Exploratory and Confirmatory Factor Analyses (EFA and CFA) were applied to the NCTE observation data in order to determine the best structure for composite scores. This data included scores from both the MQI and CLASS instruments.

For the EFA, non-orthogonal rotations (i.e., direct oblimin rotation) was used. CFA was used to account for construct-irrelevant variation caused by the use of two instruments; specifically, bi-factor models were tested to extract instrument-specific variation.

The tables from these analyses can be found in this report. The write-up of this analysis can be found in the following manuscript:

Blazar, D., Braslow, D., Charalambous, C. Y., & Hill, H. C. (submitted). *Attending to general and content-specific dimension of teaching: Exploring factors across two observation instruments.*

Exploratory Factor Analyses Loadings for a Three-Factor Solution

	Factor 1	Factor 2	Factor 3	Communalities
Eigenvalues	8.493	4.019	1.939	
Cumulative Percent of Variance Explained	32.32	46.67	52.95	
CLASS				
Negative Climate	-0.578	-0.110	-0.003	0.343
Behavior Management	0.597	0.141	0.045	0.360
Productivity	0.691	0.218	0.059	0.478
Student Engagement	0.717	0.166	-0.001	0.522
Positive Climate	0.806	0.165	0.030	0.662
Teacher Sensitivity	0.852	0.330	-0.016	0.730
Respect for Student Perspectives	0.761	0.343	0.062	0.592
Instructional Learning Formats	0.687	0.253	-0.035	0.475
Content Understanding	0.832	0.289	0.082	0.696
Analysis and Problem Solving	0.711	0.459	0.052	0.570
Quality of Feedback	0.812	0.329	0.059	0.667
Instructional Dialogue	0.841	0.410	0.031	0.729
MQI				
Linking and Connections	0.199	0.556	-0.190	0.314
Explanations	0.261	0.809	-0.236	0.657
Multiple Methods	0.119	0.549	-0.151	0.307
Generalizations	0.209	0.394	-0.098	0.162
Mathematical Language	0.352	0.363	-0.138	0.199
Remediation	0.167	0.609	-0.306	0.400
Use of Student Productions	0.332	0.889	-0.184	0.792
Student Explanations	0.236	0.808	-0.123	0.658
SMQR	0.254	0.701	-0.013	0.515
ETCA	0.296	0.839	-0.236	0.707
Major Errors	0.011	-0.195	0.835	0.698
Language Imprecisions	0.058	-0.172	0.509	0.267
Lack of Clarity	-0.005	-0.174	0.858	0.739

Notes: Extraction method is Principal Axis Factoring. Rotation method is Oblimin with Kaiser Normalization. Cells are highlighted to identify substantive factors and potential cross-loadings (i.e., loadings on two factors of similar magnitude).

Exploratory Factor Analyses Loadings for a Four-Factor Solution

	Factor 1	Factor 2	Factor 3	Factor 4	Communalities
Eigenvalues	8.493	4.019	1.939	1.479	
Cumulative Percent of Variance Explained	32.560	47.036	53.334	58.063	
CLASS					
Negative Climate	-0.459	-0.122	-0.005	-0.687	0.489
Behavior Management	0.428	0.163	0.067	0.930	0.876
Productivity	0.572	0.232	0.065	0.772	0.646
Student Engagement	0.650	0.167	-0.011	0.606	0.528
Positive Climate	0.803	0.151	0.005	0.504	0.679
Teacher Sensitivity	0.815	0.325	-0.034	0.611	0.719
Respect for Student Perspectives	0.850	0.320	0.031	0.302	0.747
Instructional Learning Formats	0.656	0.249	-0.050	0.492	0.468
Content Understanding	0.819	0.279	0.060	0.544	0.693
Analysis and Problem Solving	0.784	0.443	0.025	0.292	0.664
Quality of Feedback	0.851	0.311	0.030	0.426	0.725
Instructional Dialogue	0.896	0.392	0.000	0.416	0.811
MQI					
Linking and Connections	0.212	0.557	-0.194	0.101	0.314
Explanations	0.267	0.816	-0.238	0.158	0.671
Multiple Methods	0.162	0.546	-0.157	-0.021	0.309
Generalizations	0.198	0.398	-0.099	0.160	0.169
Mathematical Language	0.309	0.370	-0.140	0.325	0.221
Remediation	0.181	0.611	-0.308	0.075	0.401
Use of Student Productions	0.359	0.889	-0.191	0.155	0.792
Student Explanations	0.273	0.806	-0.129	0.070	0.656
SMQR	0.277	0.701	-0.018	0.114	0.516
ETCA	0.316	0.841	-0.241	0.148	0.710
Major Errors	0.018	-0.199	0.835	0.005	0.697
Language Imprecisions	0.042	-0.171	0.513	0.084	0.273
Lack of Clarity	0.006	-0.177	0.860	-0.013	0.742

Notes: Extraction method is Principal Axis Factoring. Rotation method is Oblimin with Kaiser Normalization. Cells are highlighted to identify substantive factors and potential cross-loadings (i.e., loadings on two factors of similar magnitude).

Confirmatory Factor Analysis Model Organization for Non-Bi-Factor Models

	Non-Bifactor			
Items	Model 1	Model 2	Model 3	Model 4
CLASS				
Negative Climate	Ambitious Instruction	Ambitious General Instruction	Ambitious General Instruction	Classroom Organization
Behavior Management	Ambitious Instruction	Ambitious General Instruction	Ambitious General Instruction	Classroom Organization
Productivity	Ambitious Instruction	Ambitious General Instruction	Ambitious General Instruction	Classroom Organization
Student Engagement	Ambitious Instruction	Ambitious General Instruction	Ambitious General Instruction	Ambitious General Instruction
Positive Climate	Ambitious Instruction	Ambitious General Instruction	Ambitious General Instruction	Ambitious General Instruction
Teacher Sensitivity	Ambitious Instruction	Ambitious General Instruction	Ambitious General Instruction	Ambitious General Instruction
Respect for Student Perspectives	Ambitious Instruction	Ambitious General Instruction	Ambitious General Instruction	Ambitious General Instruction
Instructional Learning Formats	Ambitious Instruction	Ambitious General Instruction	Ambitious General Instruction	Ambitious General Instruction
Content Understanding	Ambitious Instruction	Ambitious General Instruction	Ambitious General Instruction	Ambitious General Instruction
Analysis and Problem Solving	Ambitious Instruction	Ambitious General Instruction	Ambitious General Instruction	Ambitious General Instruction
Quality of Feedback	Ambitious Instruction	Ambitious General Instruction	Ambitious General Instruction	Ambitious General Instruction
Instructional Dialogue	Ambitious Instruction	Ambitious General Instruction	Ambitious General Instruction	Ambitious General Instruction
MQI				
Linking and Connections	Ambitious Instruction	Ambitious Mathematics Instruction	Ambitious Mathematics Instruction	Ambitious Mathematics Instruction
Explanations	Ambitious Instruction	Ambitious Mathematics Instruction	Ambitious Mathematics Instruction	Ambitious Mathematics Instruction
Multiple Methods	Ambitious Instruction	Ambitious Mathematics Instruction	Ambitious Mathematics Instruction	Ambitious Mathematics Instruction
Generalizations	Ambitious Instruction	Ambitious Mathematics Instruction	Ambitious Mathematics Instruction	Ambitious Mathematics Instruction
Mathematical Language	Ambitious Instruction	Ambitious Mathematics Instruction	Ambitious Mathematics Instruction	Ambitious Mathematics Instruction
Remediation	Ambitious Instruction	Ambitious Mathematics Instruction	Ambitious Mathematics Instruction	Ambitious Mathematics Instruction
Use Productions	Ambitious Instruction	Ambitious Mathematics Instruction	Ambitious Mathematics Instruction	Ambitious Mathematics Instruction
Student Explanations	Ambitious Instruction	Ambitious Mathematics Instruction	Ambitious Mathematics Instruction	Ambitious Mathematics Instruction
SMQR	Ambitious Instruction	Ambitious Mathematics Instruction	Ambitious Mathematics Instruction	Ambitious Mathematics Instruction
ETCA	Ambitious Instruction	Ambitious Mathematics Instruction	Ambitious Mathematics Instruction	Ambitious Mathematics Instruction
Major Errors	Ambitious Instruction	Ambitious Mathematics Instruction	Mathematical Errors	Mathematical Errors
Language Imprecisions	Ambitious Instruction	Ambitious Mathematics Instruction	Mathematical Errors	Mathematical Errors
Lack of Clarity	Ambitious Instruction	Ambitious Mathematics Instruction	Mathematical Errors	Mathematical Errors
Number of Factors	1	2	3	4
Nested in	M2-M8	M3-M8	M4	

Confirmatory Factor Analysis Model Organization for Bi-Factor Models

Items	Model 5	Model 6	Model 7	Model 8
CLASS				
Negative Climate	Ambitious Instruction	Ambitious Instruction	Classroom Organization	Classroom Organization
Behavior Management	Ambitious Instruction	Ambitious Instruction	Classroom Organization	Classroom Organization
Productivity	Ambitious Instruction	Ambitious Instruction	Classroom Organization	Classroom Organization
Student Engagement	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction
Positive Climate	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction
Teacher Sensitivity	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction
Respect for Student Perspectives	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction
Instructional Learning Formats	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction
Content Understanding	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction
Analysis and Problem Solving	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction
Quality of Feedback	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction
Instructional Dialogue	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction
MQI				
Linking and Connections	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction
Explanations	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction
Multiple Methods	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction
Generalizations	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction
Mathematical Language	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction
Remediation	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction
Use Productions	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction
Student Explanations	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction
SMQR	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction
ETCA	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction	Ambitious Instruction
Major Errors	Ambitious Instruction	Mathematical Errors	Ambitious Instruction	Mathematical Errors
Language Imprecisions	Ambitious Instruction	Mathematical Errors	Ambitious Instruction	Mathematical Errors
Lack of Clarity	Ambitious Instruction	Mathematical Errors	Ambitious Instruction	Mathematical Errors
Number of Factors	3	4	4	5
Nested in	M6-M8	M8	M8	

Note: All models also include two method factors with all items cross loading onto their respective instrument factors.

Standardized Factor Loadings for CFA Model 4

Items	Ambitious Mathematics Instruction	Ambitious General Instruction	Classroom Organization	Mathematical Errors
CLASS				
Negative Climate			0.699***	
Behavior Management			-0.841***	
Productivity			-0.883***	
Student Engagement	0.671***			
Positive Climate	0.797***			
Teacher Sensitivity	0.823***			
Respect for Student Perspectives	0.821***			
Instructional Learning Formats	0.673***			
Content Understanding	0.831***			
Analysis and Problem Solving	0.780***			
Quality of Feedback	0.856***			
Instructional Dialogue	0.886***			
MQI				
Linking and Connections		0.524***		
Explanations		0.759***		
Multiple Methods		0.523***		
Generalizations		0.389***		
Mathematical Language		0.368***		
Remediation		0.575***		
Use Productions		0.909***		
Student Explanations		0.836***		
SMQR		0.746***		
ETCA		0.848***		
Major Errors				0.834***
Language Imprecisions				0.508***
Lack of Clarity				0.876***

Notes: ~ $p<0.10$, * $p<0.05$, ** $p<0.01$, *** $p<0.001$

Standardized Factor Loadings for CFA Model 8

	Instrument Factors		Substantive Factors		
Items	CLASS	MQI	Ambitious Instruction	Classroom Organization	Mathematical Errors
CLASS					
Negative Climate	-0.493***			-0.486***	
Behavior Management	0.451***			0.836***	
Productivity	0.619***			0.559***	
Student Engagement	0.619***		0.237**		
Positive Climate	0.808***		0.100~		
Teacher Sensitivity	0.795***		0.201**		
Respect for Student Perspectives	0.756***		0.333***		
Instructional Learning Formats	0.615***		0.266***		
Content Understanding	0.855***		0.078		
Analysis and Problem Solving	0.719***		0.326***		
Quality of Feedback	0.849***		0.180**		
Instructional Dialogue	0.820***		0.348***		
MQI					
Linking and Connections		0.573***	0.137		
Explanations		0.903***	0.133		
Multiple Methods		0.486***	0.245~		
Generalizations		0.428***	0.084		
Mathematical Language		0.382***	0.113		
Remediation		0.704***	0.050		
Use Productions		0.604**	0.722***		
Student Explanations		0.617***	0.578**		
SMQR		0.432*	0.654***		
ETCA		0.636**	0.535*		
Major Errors		-0.238***			-0.788***
Language Imprecisions		-0.166**			-0.477***
Lack of Clarity		-0.197**			-0.870***

Notes: ~ p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Section 2.c.i MQI – Reliability Analyses

Item-level Reliability

code	segments	% agreement	% agreement within 1	kappa
etca	12418	0.67	0.98	0.24
expl	12427	0.70	0.98	0.23
langimp	12427	0.80	0.99	0.25
lcp	12428	0.86	0.99	0.18
link	12426	0.70	0.97	0.31
majerr	12426	0.91	0.99	0.24
mgen	12431	0.95	0.99	0.15
mlang	12416	0.59	0.98	0.23
mmeth	12434	0.85	0.99	0.42
remed	12417	0.66	0.96	0.27
smqr	12423	0.76	0.98	0.25
stexpl	12425	0.80	0.99	0.36
useprod	12420	0.76	0.98	0.25

code	Y1 ICC	Y1 Adjusted ICC	Y2 ICC	Y2 Adjusted ICC	Y3 ICC	Y3 Adjusted ICC	ICC	Adjusted ICC
etca	0.28	0.54	0.16	0.36	0.19	0.42	0.21	0.62
expl	0.14	0.32	0.10	0.26	0.11	0.26	0.17	0.55
langimp	0.17	0.38	0.15	0.35	0.13	0.30	0.11	0.43
lcp	0.19	0.41	0.08	0.20	0.06	0.15	0.14	0.49
link	0.16	0.37	0.11	0.28	0.14	0.34	0.15	0.51
majerr	0.11	0.26	0.08	0.20	0.07	0.19	0.10	0.41
mgen	0.00	0.00	0.00	0.00	0.05	0.14	0.04	0.20
mlang	0.12	0.28	0.08	0.21	0.05	0.14	0.08	0.34
mmeth	0.19	0.42	0.14	0.33	0.01	0.04	0.15	0.52
remed	0.19	0.42	0.16	0.36	0.05	0.14	0.16	0.53
smqr	0.24	0.49	0.18	0.39	0.17	0.38	0.19	0.59
stexpl	0.30	0.56	0.32	0.58	0.19	0.42	0.30	0.72
useprod	0.22	0.45	0.33	0.59	0.21	0.45	0.24	0.66

Procedure

- Keep NCTE videos in Year 1, Year 2, and Year 3
- In Year 3, the *Use Mathematical Contributions* code was recoded to reflect a *Low, Mid,* and *High* scoring point system comparable to that of the *Use Mathematical Productions* code in Year 1 and Year 2
- Video-rater sample
    - Dropped if < 2 raters
    - If > 2 raters, randomly selected 2
- *Segments* variable indicates segments within chapters with scores from both raters for the item

Reliability Measures

- *% agreement* – indicates percentage of segments where both raters agreed on score for the item
- *% agreement within 1* – indicates percentage of segments where both raters scored the segment the same score, or within one score point (i.e., low and mid, mid and high) for the item
- *Kappa* – Cohen's kappa
- *ICC*
    - Collapse scores for each item across raters and segments to the lesson level
    - Estimate the following multilevel model, where lesson *l* scores for each item are nested within teachers *k*:
        - $\text{MQI}_{lk} = \beta_0 + \mu_k + \varepsilon_{lk}$
    - Calculate the ICC using the following equation:
        - $\text{ICC} = \frac{\text{var}(\mu_k)}{\text{var}(\mu_k)+\text{var}(\varepsilon_{lk})}$
- *Adjusted ICC*
    - Collapse scores for each item across raters and segments to the lesson level
    - Estimate the following multilevel model, where lesson *l* scores for each item are nested within teachers *k*:
        - $\text{MQI}_{lk} = \beta_0 + \mu_k + \varepsilon_{lk}$
    - Calculate the Adjusted ICC using the following equation:
        - $\text{ADJICC} = \frac{\text{var}(\mu_k)}{\text{var}(\mu_k)+\frac{\text{var}(\varepsilon_{lk})}{n_l}}$
    - n_l
        - Within-year ICCs: 3
        - Overall: 6

Dimension-level Reliability

dimension	items
Ambitious Instruction	link, expl, mmeth, mlang, mgen, remed, useprod, stexpl, smqr, etca
Errors	majerr, langimp, lcp
Richness – 3 Factors	link, expl, mmeth, mlang, mgen, remed
Richness – 4 Factors	link, expl, mmeth, mlang, mgen
SPMMR – 3 Factors	useprod, stexpl, smqr, etca
SPMMR – 4 Factors	stexpl, smqr, etca
Working with Students – 4 Factors	useprod, remed

dimension	segments	% agreement	% agreement within 1	kappa
Ambitious Instruction	124237	0.74	0.98	0.31
Errors	37281	0.86	0.99	0.24
Richness – 3 Factors	74551	0.74	0.98	0.33
Richness – 4 Factors	62134	0.76	0.98	0.34
SPMMR – 3 Factors	49686	0.75	0.98	0.28
SPMMR – 4 Factors	37266	0.74	0.98	0.29
Working with Students – 4 Factors	24837	0.71	0.97	0.27

dimension	Y1 ICC	Y1 Adjusted ICC	Y2 ICC	Y2 Adjusted ICC	Y3 ICC	Y3 Adjusted ICC	ICC	Adjusted ICC	Cronbach's Alpha
Ambitious Instruction	0.33	0.59	0.25	0.50	0.21	0.45	0.32	0.74	0.70
Errors	0.19	0.42	0.17	0.37	0.17	0.38	0.18	0.56	0.45
Richness – 3 Factors	0.27	0.52	0.15	0.35	0.12	0.29	0.25	0.66	0.51
Richness – 4 Factors	0.26	0.51	0.14	0.32	0.11	0.27	0.22	0.63	0.49
SPMMR – 3 Factors	0.35	0.62	0.34	0.60	0.28	0.54	0.33	0.74	0.68
SPMMR – 4 Factors	0.36	0.62	0.31	0.57	0.26	0.51	0.32	0.74	0.56
Working with Students – 4 Factors	0.24	0.49	0.27	0.53	0.15	0.34	0.24	0.65	0.26

Procedure

- Keep NCTE videos in Year 1, Year 2, and Year 3
- In Year 3, the *Use Mathematical Contributions* code was recoded to reflect a *Low, Mid,* and *High* scoring point system comparable to that of the *Use Mathematical Productions* code in Year 1 and Year 2
- Video-rater sample
    - Dropped if < 2 raters
    - If > 2 raters, randomly selected 2
- *Segments* variable indicates segments within chapters with scores from both raters for the **all items in the dimension** (thus, resulting in greater number of segments for the *Ambitious Instruction* dimension, which comprises ten items)

Reliability Measures

- *% agreement* – indicates percentage of segments where both raters agreed on score for items in the dimension
- *% agreement within 1* – indicates percentage of segments where both raters scored the segment the same score, or within one score point (i.e., low and mid, mid and high) for the items in the dimension
- *Kappa* – Cohen's kappa
- *ICC*
    - Collapse scores for each item across raters and segments to the lesson level
    - Take averages of lesson-level item scores to create lesson-level dimension scores
    - Estimate the following multilevel model, where lesson *l* scores for each dimension are nested within teachers *k*:
        - $\mathrm{MQI}_{lk} = \beta_0 + \mu_k + \varepsilon_{lk}$
    - Calculate the ICC using the following equation:
        - $\mathrm{ICC} = \frac{\mathrm{var}(\mu_k)}{\mathrm{var}(\mu_k)+\mathrm{var}(\varepsilon_{lk})}$
- *Adjusted ICC*
    - Collapse scores for each item across raters and segments to the lesson level
    - Take averages of lesson-level item scores to create lesson-level dimension scores
    - Estimate the following multilevel model, where lesson *l* scores for each dimension are nested within teachers *k*:
        - $\mathrm{MQI}_{lk} = \beta_0 + \mu_k + \varepsilon_{lk}$
    - Calculate the Adjusted ICC using the following equation:
        - $\mathrm{ADJICC} = \frac{\mathrm{var}(\mu_k)}{\mathrm{var}(\mu_k)+\frac{\mathrm{var}(\varepsilon_{lk})}{n_l}}$
    - n_l
        - Within-year ICCs: 3
        - Overall: 6
- *Cronbach's Alpha*
    - Collapse scores for each item across raters and segments to the lesson level
    - Calculate alpha of all lesson-level scores for all items within a dimension

Section 2.c.ii CLASS – Reliability Analyses

Item-level Reliability

code	Y1 ICC	Y1 Adjusted ICC	Y2 ICC	Y2 Adjusted ICC	Y3 ICC	Y3 Adjusted ICC	ICC	Adjusted ICC
claps	0.03	0.08	0.06	0.17	0.05	0.14	0.06	0.27
clbm	0.35	0.62	0.17	0.37	0.11	0.27	0.24	0.65
clcu	0.03	0.07	0.00	0.00	0.00	0.00	0.01	0.06
clilf	0.12	0.28	0.10	0.24	0.07	0.19	0.08	0.33
clinstd	0.11	0.26	0.12	0.29	0.00	0.00	0.08	0.36
clnc	0.11	0.27	0.16	0.36	0.01	0.04	0.14	0.50
clpc	0.21	0.45	0.15	0.35	0.12	0.29	0.18	0.56
clprdt	0.06	0.17	0.08	0.22	0.00	0.01	0.06	0.29
clqf	0.00	0.00	0.07	0.18	0.05	0.13	0.06	0.26
clrsp	0.14	0.32	0.09	0.23	0.04	0.12	0.09	0.36
clsteng	0.07	0.18	0.06	0.16	0.00	0.00	0.06	0.28
clts	0.11	0.26	0.13	0.31	0.02	0.05	0.10	0.39

Procedure

- Keep NCTE videos in Year 1, Year 2, and Year 3
- Reverse score *clnc* (Negative Climate)
- Video-rater sample
    - If > 1 raters, randomly selected 1

Reliability Measures

- *ICC*
    - Collapse scores for each item across segments to the lesson level
    - Estimate the following multilevel model, where lesson *l* scores for each item are nested within teachers *k*:
        - $\text{CLASS}_{lk} = \beta_0 + \mu_k + \varepsilon_{lk}$
    - Calculate the ICC using the following equation:
        - $\text{ICC} = \frac{\text{var}(\mu_k)}{\text{var}(\mu_k)+\text{var}(\varepsilon_{lk})}$
- *Adjusted ICC*
    - Collapse scores for each item across segments to the lesson level
    - Estimate the following multilevel model, where lesson *l* scores for each item are nested within teachers *k*:
        - $\text{CLASS}_{lk} = \beta_0 + \mu_k + \varepsilon_{lk}$
    - Calculate the Adjusted ICC using the following equation:
        - $\text{ADJICC} = \frac{\text{var}(\mu_k)}{\text{var}(\mu_k)+\frac{\text{var}(\varepsilon_{lk})}{n_l}}$
    - n_l
        - Within-year ICCs: 3
        - Overall: 6

Dimension-level Reliability

dimension	items
Classroom Organization	clbm, clnc, clprdt
Support	clpc, clts, clrsp, clilf, clcu, claps, clqf, clinstd, clsteng
Emotional Support	clpc, clts, clrsp
Instructional Support	clilf, clcu, claps, clqf, clinstd, clsteng

dimension	Y1 ICC	Y1 Adjusted ICC	Y2 ICC	Y2 Adjusted ICC	Y3 ICC	Y3 Adjusted ICC	ICC	Adjusted ICC	Cronbach's Alpha
Classroom Organization	0.34	0.61	0.18	0.40	0.05	0.13	0.22	0.63	0.72
Support	0.18	0.40	0.13	0.30	0.04	0.12	0.13	0.47	0.90
Emotional Support	0.20	0.43	0.16	0.36	0.10	0.25	0.16	0.53	0.80
Instructional Support	0.12	0.29	0.09	0.22	0.01	0.03	0.09	0.36	0.87

Procedure

- Keep NCTE videos in Year 1, Year 2, and Year 3
- Reverse score *clnc* (Negative Climate)
- Video-rater sample
    - If > 1 raters, randomly selected 1

Reliability Measures

- *ICC*
    - Collapse scores for each item across segments to the lesson level
    - Take averages of lesson-level item scores to create lesson-level dimension scores
    - Estimate the following multilevel model, where lesson *l* scores for each dimension are nested within teachers *k*:
        - $\text{CLASS}_{lk} = \beta_0 + \mu_k + \varepsilon_{lk}$
    - Calculate the ICC using the following equation:
        - $\text{ICC} = \frac{\text{var}(\mu_k)}{\text{var}(\mu_k)+\text{var}(\varepsilon_{lk})}$
- *Adjusted ICC*
    - Collapse scores for each item across segments to the lesson level
    - Take averages of lesson-level item scores to create lesson-level dimension scores
    - Estimate the following multilevel model, where lesson *l* scores for each dimension are nested within teachers *k*:
        - $\text{CLASS}_{lk} = \beta_0 + \mu_k + \varepsilon_{lk}$
    - Calculate the Adjusted ICC using the following equation:
        - $\text{ADJICC} = \frac{\text{var}(\mu_k)}{\text{var}(\mu_k)+\frac{\text{var}(\varepsilon_{lk})}{n_l}}$
    - n_l
        - Within-year ICCs: 3
        - Overall: 6
- *Cronbach's Alpha*
    - Collapse scores for each item across segments to the lesson level
    - Calculate alpha of all lesson-level scores for all items within a dimension

Section 2.d. MQI – Generalizability Study Results

Note: results from the following analysis provide technical information on the MQI obtained from prior work on **non-NCTE** observation data.

Problem of the study. Recent years have seen significant emphasis on obtaining reliable estimates of teaching quality yielded from classroom observations. In exploring the reliability of such estimates, researchers have for years focused on inter-rater reliabilities (see, for example, Heneman & Milanowski, 2003; Sartain, Stoelinga, & Brown, 2009). Recognizing that inter-rater reliability is not enough, we sought to examine different elements in an observational system that might contribute to the reliability of the estimates obtained through classroom observations. By *observational systems* we defined a set of elements that collectively contribute to producing scores representing individual teachers' instructional quality (Hill, Charalambous, & Kraft, 2012). These include, for instance, the observational instrument itself, the raters conducting the observations (as well as rater training and certification), the number and length of observations to be collected per teacher, the period during the year in which the observations are conducted, and the number of raters per observation.

Research questions. To empirically examine how certain elements of an observational system contribute to the reliability of the teaching-quality estimates obtained, we capitalized on the Generalizability framework (Brennan, 2001; Marcoulides, 1989; Shavelson & Webb, 1991) and our work in developing and refining the Mathematical Quality of Instruction (MQI) instrument. Our investigation was guided by the following research questions:

- How different elements within a given observational system (e.g., number of lessons to be observed, length of observation, number of raters conducting the observations, and teaching dimensions under consideration) might affect the reliability of the estimates obtained when using MQI?
- What is the optimal number of raters needed per observation and the number of lessons needed per teacher to obtain acceptably reliable estimates of instructional quality when using the MQI?

Methods. To address these questions we focused on three dimensions of the MQI instrument: richness of the mathematics, errors and imprecisions, and student participation in meaning making and reasoning. These dimensions, as well as their respective codes, were used to code 24 lessons offered by eight middle-grade teachers with different levels of mathematical knowledge (each teaching three lessons). We also recruited 10 graduate students and former teachers who received a two-day intensive training; at the culmination of the training all took a certification exam. The nine raters who passed the certification exam were asked to code the 24 lessons using all the codes within each of the three MQI dimension, as well as a holistic code per dimension (this latter code represented a holistic evaluation of the observed lesson segment and was not necessarily identical to the overall/average score that each segment would receive based on the codes falling in each dimension). Each rater assigned scores of low, medium, and high for all the

items under consideration for every single 7.5-min lesson segment (typically lessons included six to eight such segments). To analyze the data, we first aggregated the segment scores to the lesson level, acknowledging that it would not be reasonable to expect manifestations of each single code within each lesson segment. Using these data and employing a two-facet design with lessons nested with teachers and crossed with raters, we then conducted a Generalizability-study (G-study) to determine the variance components that could be attributed to teachers, lessons, and raters, as well as their two-way interactions, and the combination of the three-way interaction and the measurement error. We calculated these variance components for each of the individual MQI items. We also aggregated across items, and calculated an "average" estimate for each dimension, in addition to the holistic code referred to above. Using the variance decomposition yielded from the G-study, we then conducted a series of Design studies (D-studies) to determine the optimal number of lessons and raters needed to code each lesson to obtain reliable estimates of teaching quality.

Results. The main results emerging from this exploration are summarized below:

- There were notable differences in the variance decomposition across the MQI items. For example, for Richness, two items (representations and developing generalizations) exhibited negligible teacher-level variation, whereas a large portion of the variance lied at the lesson level; in contrast, for other codes of this dimension (explanations, multiple solutions, and mathematical language) a significant portion of the variance (from 22% to 42%) was associated with the teacher level (see Table 1 in the Appendix). Interestingly, reporting inter-rater agreement percentages would yield a totally different picture, since, for instance, representations and mathematical language had similar agreement rates (69% and 55%, respectively).
- For all three dimensions, the "average" code, as well as the "holistic" code exhibited higher percentages of teacher-level variance compared to the individual codes per dimension (see Tables 2 and 3 in the Appendix). This implied that these codes could yield more reliable estimates of teaching quality, compared to the individual codes per each dimension.
- A combination of three lessons coded by two raters each yielded sufficiently reliable estimates of teaching quality for all three MQI dimensions (0.77 for Richness, 0.71 for Errors and Imprecisions, and 0.81 for Student participation in meaning making and reasoning).Adding a fourth lesson or a third rater increased reliabilities only minimally (see Figures 1-3 in the Appendix).
- We also explored if reliability estimates would change if we observed only the first 30 minutes per lesson rather than the entire lesson period. This is typically the case in which school principals or superintendents might find themselves when pressed for time and when rushing to observe many teachers in a short period of time. Our investigation suggested that this modification would yield only minimal differences for two of the MQI dimensions: Richness and Errors and Imprecisions; however, it would reduce reliability significantly for the dimension of Student Participation in Meaning making and reasoning. Specifically, whereas the reliability estimates for the optimal combination of three-lessons/two raters would largely remain unchanged for the first two dimensions, it would drop per .13 in the last dimension (see Table 4 in the Appendix).

- All these results were based on the relative reliability coefficient (rho), which is associated with making relative decisions (e.g., rank ordering teachers). When high-stakes decisions are expected to be made (as is the case with hiring or firing teachers), the absolute coefficient (phi) is needed. In this latter case, our investigation suggested that more raters per lesson and lessons per teacher would be needed to be observed (i.e., at least one additional rater would be needed for the first two dimensions; three additional raters per lesson and three additional lessons per teacher would be needed for the last dimension).

In sum, our findings point to the importance of closely examining how different elements within a given observational system affect the reliability of the estimates obtained from classroom observations.

References:

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer-Verlag.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The *dependability of behavioral measurements: Theory of generalizability scores and profiles*. New York: John Wiley.

Heneman, H. G., & Milanowski, A. T. (2003). Continuing assessment of teacher reactions to a Standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education 17*, 173-195.

Hill, H. C., Charalambous, C. Y., & Kraft, M. (2012). When rater reliability is not enough: Teacher observation systems and a case for the G-study. *Educational Researcher, 41*(2), 56-64.

Marcoulides, G. A. (1989). The application of generalizability analysis to observational studies. *Quality and Quantity, 23*, 115-127.

Sartain, L., Stoelinga, S. R., & Brown, E. (2009). *Evaluation of the excellent in teaching pilot: A report to the Joyce Foundation*. Chicago: Consortium on Chicago School Research.

Shavelson. R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: Sage.

Appendix

Table 1.
Variance Decomposition for the Richness *Dimension of the* Mathematical Quality of Instruction *Instrument*

	Richness of the Mathematics						
	Individual Items					Overall Richness	
	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)
	Representations	Multiple Solution Procedures/ Solutions	Explanations	Developing Generalizations	Mathematical Language	Average of Items (I)-(V)	Holistic
Teachers (t)	0.97	34.61	22.01	0.00	41.55	42.52	45.70
Lessons: teacher (l:t)	20.24	28.00	12.01	23.94	16.99	10.52	2.76
Raters (r)	9.14	2.61	21.48	7.61	4.99	6.17	13.96
Teachers*Raters (t*r)	0.00	0.00	8.52	13.12	3.64	7.83	3.27
Residual ((l:t)*r, e)	69.65	34.78	35.99	55.33	32.83	32.97	34.31
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Note: Cells represent the percent of variance explained by different facets in a G-study

Table 2
Variance Decomposition for the Errors and Imprecision *Dimension of the* Mathematical Quality of Instruction *Instrument*

	Errors and Imprecision				
	Individual Items			Overall Errors and Imprecisions	
	(I)	(II)	(III)	(IV)	(V)
	Major Errors	Notation & Language	Lack of Clarity	Average of Items (I)-(III)	Holistic
Teachers (t)	13.11	31.23	21.12	31.88	36.04
Lessons: teacher (l:t)	5.90	7.22	11.71	8.81	3.20
Raters (r)	2.59	10.72	9.53	13.04	13.51
Teachers*Raters (t*r)	14.35	5.62	1.61	6.45	5.26
Residual ((l:t)*r, e)	64.04	45.21	56.03	39.82	41.99
Total	100.00	100.00	100.00	100.00	100.00

Note: Cells represent the percent of variance explained by different facets in a G-study

Table 3
Variance Decomposition for the Student Participation in Meaning Making and Reasoning *Dimension of the* Mathematical Quality of Instruction *Instrument*

	Student Participation in Mathematical Meaning-Making and Reasoning				
	Individual Items			Overall SPMMR	
	(I)	(II)	(III)	(IV)	(V)
	Student Explanations	Student Questioning & Reasoning	Enacted Task Demand	Average of Items (I)-(III)	Holistic
Teachers (t)	17.77	14.16	21.96	32.78	27.11
Lessons: teacher (l:t)	39.81	11.74	6.09	7.22	2.09
Raters (r)	10.71	33.10	24.19	28.58	27.12
Teachers*Raters (t*r)	2.26	0.05	1.23	0.00	2.48
Residual ((l:t)*r, e)	29.45	40.94	46.52	31.43	41.19
Total	100.00	100.00	100.00	100.00	100.00

Note: Cells represent the percent of variance explained by different facets in a G-study

Table 4

Comparison of the Reliability Estimates (ρ) for Different Combinations of Raters and Lessons for the Whole Lesson and the First Thirty Minutes of a Lesson

	Richness		Errors and Imprecision		Student Participation in Meaning Making and Reasoning	
	Whole lesson	30 minutes	Whole lesson	30 minutes	Whole lesson	30 minutes
One lesson						
1 Rater	0.45	0.50	0.37	0.34	0.46	0.32
2 Raters	0.58	0.59	0.50	0.46	0.59	0.41
3 Raters	0.64	0.63	0.57	0.53	0.65	0.45
4 Raters	0.67	0.65	0.61	0.57	0.68	0.48
Two lessons						
1 Rater	0.59	0.65	0.51	0.49	0.63	0.49
2 Raters	0.71	0.73	0.64	0.62	0.74	0.58
3 Raters	0.76	0.77	0.71	0.68	0.79	0.62
4 Raters	0.79	0.78	0.74	0.71	0.81	0.65
Three lessons						
1 Rater	0.66	0.73	0.58	0.57	0.72	0.59
2 Raters	0.77	0.80	0.71	0.70	0.81	0.68
3 Raters	0.81	0.83	0.77	0.75	0.85	0.71
4 Raters	0.84	0.84	0.80	0.78	0.87	0.73
Four lessons						
1 Rater	0.69	0.77	0.63	0.63	0.77	0.66
2 Raters	0.80	0.83	0.75	0.74	0.85	0.74
3 Raters	0.84	0.86	0.81	0.79	0.88	0.77
4 Raters	0.86	0.87	0.83	0.82	0.90	0.78

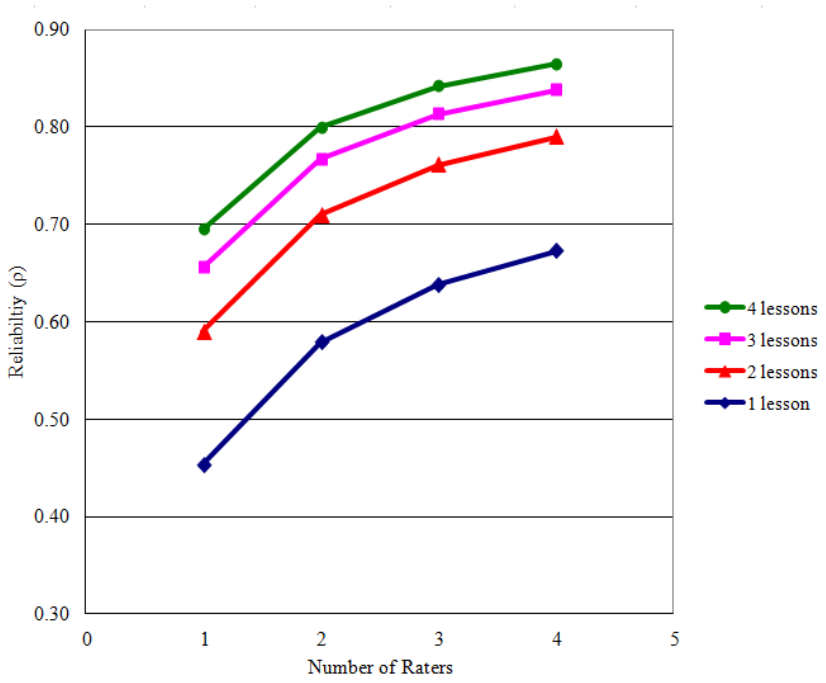


Figure 1. The reliability of different combinations of raters and lessons for *Richness*.

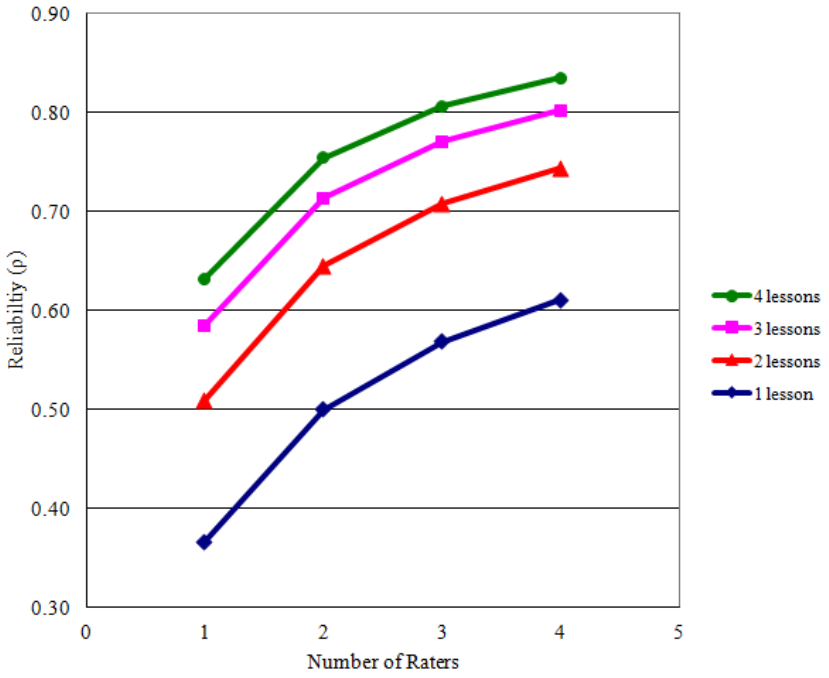


Figure 2: The reliability of different combinations of raters and lessons for *Errors and Imprecisions*.

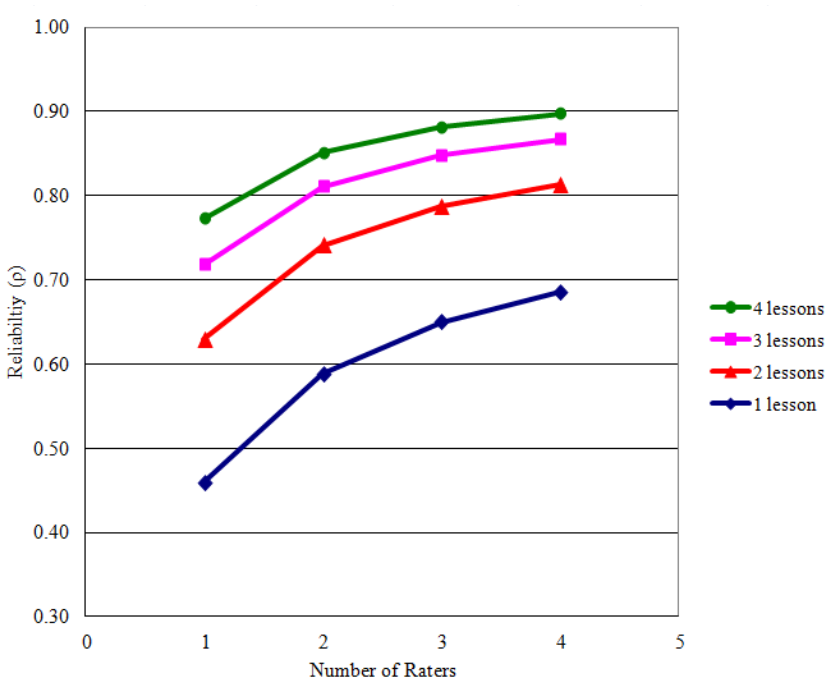


Figure 3: The reliability of different combinations of raters and lessons for *Student participation in mathematical meaning-making and reasoning.*

Section 3. Other NCTE Measures

a. Study-developed Student Assessment

Students in participating NCTE classrooms responded to an additional mathematics assessment whose items were designed to be more aligned with the various teacher effectiveness measures collected throughout the study. The test, developed in partnership with ETS, was administered in the fall and spring semesters of each academic year. More detail about the properties of the assessment can be found in the technical report (Hickman, Fu, & Hill, 2012), also distributed with this data deposit.

b. Tripod Student Survey

Students in participating NCTE classrooms also responded to a student survey administered in the spring semesters of each academic year. Certain items on the surveys were adapted from the Tripod student survey, with the permission of the Tripod Project. For more information on the Tripod, contact Rob@TripodEd.com or visit www.TripodEd.com.