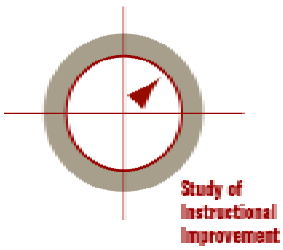

*Measuring Teachers' Pedagogical Content Knowledge in Surveys: An Exploratory Study**

Brian Rowan
Steven G. Schilling
Deborah L. Ball
Robert Miller

With

Sally Atkins-Burnett
Eric Camburn
Delena Harrison
Geoff Phelps

October, 2001

D R A F T

* Work on this paper was supported by grants to the Consortium for Policy Research in Education from the Educational Statistics Services Institute of the American Institutes for Research, the Atlantic Philanthropies –North America, the Office of Educational Research and Improvement of the U.S. Department of Education, and the National Science Foundation. The opinions expressed here are those of the authors and are not specifically endorsed by the sponsors.

Measuring Teachers' Pedagogical Content Knowledge in Surveys: An Exploratory Study

This paper discusses the efforts of a group of researchers at the University of Michigan to develop survey-based measures of what Lee S. Shulman (1986; 1987) called teachers' "pedagogical content knowledge." In the paper, we briefly discuss our rationale for using a survey instrument to measure teachers' pedagogical content knowledge and report on the results of a pilot study in which a bank of survey items was developed to directly measure this construct in two critical domains of the elementary school curriculum—reading/ language arts and mathematics. In the paper, we demonstrate that particular facets of teachers' pedagogical content knowledge can be measured reliably with as few as 6 – 10 survey items. However, we point to additional methodological and conceptual issues that must be addressed if sound, survey-based measures of teachers' pedagogical content knowledge are to be developed for use in large-scale, survey-based research on teaching.

Rationale

Most observers agree that successful teachers draw on specialized knowledge in their instructional work with students, but specifying and measuring this knowledge has proven elusive and controversial in American education. One particular issue that has clouded efforts to conceptualize and measure the knowledge base for teaching has been the perceived distinction between teachers' subject matter knowledge and teachers' knowledge of general pedagogical principles and practices. During most of the 20th century, this distinction has been reified in a variety of bureaucratic and institutional arrangements in American education. For example, teacher preparation programs have been organized under the assumption that prospective teachers will acquire subject matter knowledge in courses taken in the arts and sciences, but that they will acquire knowledge of pedagogy in *separate* classes taken in education schools. Similarly, separate licensing examinations have been developed in American education, some designed to test teachers' subject matter knowledge, and others to test teachers' knowledge of general pedagogical principles and practices. Even teacher assessment practices and associated research on teaching in the United States have tended to maintain a distinction between teachers' knowledge of subject-matter content and teachers' knowledge of pedagogy. For example, research and evaluation efforts frequently try to measure teachers' use of a general set of pedagogical practices under the assumption that these practices are instructionally effective no matter what the academic subject or grade level being taught and without regard for the knowledge that teachers have of the academic content they are teaching.

Since the 1980's, however, the analytic distinction between teachers' subject matter knowledge and teachers' knowledge of pedagogy has begun to fade, in large part due to Lee Shulman's (1986; 1987) seminal work on teachers' "pedagogical content knowledge." Shulman argued that a distinctive form of teachers' professional knowledge that he called pedagogical content knowledge exists, and that this form of knowledge builds upon, but is different from, teachers' subject matter knowledge or knowledge of general principles of pedagogy. In Shulman's view, pedagogical content knowledge is a form of *practical* knowledge that is used by teachers to guide their actions in highly contextualized classroom settings. In Shulman's view, this form of practical knowledge entails, among other things: (a)

knowledge of how to structure and represent academic content for direct teaching to students; (b) knowledge of the common conceptions, misconceptions, and difficulties that students encounter when learning particular content; and (c) knowledge of the specific teaching strategies that can be used to address students' learning needs in particular classroom circumstances. In the view of Shulman (and others), pedagogical content knowledge builds on other forms of professional knowledge, and is therefore a critical—and perhaps even the paramount—constitutive element in the knowledge base of teaching.

The Problem

Shulman's (1986; 1987) ideas have had an important impact on American education. Since they were first published, mainstream research on teaching has increasingly moved beyond the search for pedagogical principles that can be generalized across grade levels and academic subjects and toward inquiries into the specific forms of pedagogical and content knowledge that teachers bring to bear when teaching particular academic content to students at particular grade levels. In line with this new conception of the knowledge base for teaching, efforts are being made to bring about a closer integration of academic and professional coursework in teacher education programs. Moreover, new conceptions of teachers' professional knowledge are affecting teacher assessment practices and licensing examinations in education. The widely-used Praxis series, published by the Educational Testing Service, for example, has been revised to measure not only the subject matter knowledge of prospective teachers, but also their pedagogical content knowledge in specific areas of the school curriculum.

Unfortunately, none of these trends are well-reflected in contemporary *survey* research on teaching. To be sure, survey researchers in education are interested in measuring teachers' professional knowledge and in correlating such knowledge to patterns of student achievement in schools. But survey research in this area has tended to measure teachers' knowledge indirectly or in ways that only loosely intersect with emerging views of pedagogical content knowledge and its role in teacher performance. For example, an extensive line of survey research, dating to the Coleman report (Coleman et al., 1966), shows that teachers' general cognitive ability—as assessed by teachers' scores on verbal ability tests, basic skills tests, and college entrance examinations—is significantly correlated to patterns of student achievement in schools (see, for example, the meta-analysis reported in Greenwald, Hedges, and Laine, 1996 and the review of more recent work in this area by Ferguson and Brown, 2000). But decades of research in personnel psychology, as well as research on assessment practices in education, shows that measures of general cognitive ability are among the weakest predictors of job performance (Prediger, 1989; Smith & George, 1992; Porter, Youngs, and Odden, in press). As a result, personnel psychologists now advise moving beyond the use of measures of general cognitive ability and to instead examine the role of job-relevant knowledge in predicting job performance.

To their credit, survey researchers in the field of education have moved in this direction, especially in studies examining the effects of teachers' subject matter knowledge and/or knowledge of pedagogy on students' academic achievements. But survey research on this topic has been limited by available measures. For example, several large-scale studies (reviewed in Rowan et al., 1997 and Brewer and Goldhaber, 2000) have tried to assess the effect of *teachers' subject matter knowledge* on students' achievement by examining differences in

student outcomes for teachers with different academic majors. In general, these studies have been conducted in high schools and have shown that in classes where teachers have an academic major in the subject area in which students are being tested, the tested students have higher adjusted achievement gains. The effect sizes reported in this literature are quite small, however. In *NELS: 88* data, for example, the *r*-type effect sizes for such variables were .05 for science gains, and .01 for math gains.[1] Other large-scale studies of teaching have tried to examine the effects on student achievement associated with teachers' *pedagogical knowledge*. A study by Monk (1994) is particularly noteworthy in this regard, showing that the number of classes in subject-matter pedagogy taken by teachers' during their college years had positive effects on high school students' adjusted achievement gains. Darling-Hammond and colleagues (1995) cite additional, small-scale studies supporting this conclusion. But once again, the effect sizes reported in this literature are quite small, perhaps because the measures used in the research are proxy measures of the key constructs of interest.

All of this suggests that survey researchers face an important problem in their efforts to measure teachers' professional knowledge and in analyzing its effects on student achievement. Current conceptions of the knowledge base for teaching are changing, and assessment practices in other sectors of the education community are changing to reflect this fact. Survey research in education, by contrast, continues to rely on measures of teachers' general cognitive ability or indirect and unsatisfactory proxy measures of teachers' job-relevant knowledge to assess the effects of teachers' knowledge on student achievement in schools.

Procedures

In order to address this problem, we have been developing a set of questionnaire items designed to measure teachers' "pedagogical content knowledge" within the context of a multi-purpose survey being conducted in elementary schools participating in three of America's largest comprehensive school reform programs.[2] As far as we can tell, our efforts in this area are relatively pioneering. For example, a review of the literature found only one similar effort in this area. As part of the Teaching and Learning to Teach (TELT) study conducted at Michigan State University, researchers set out to develop a battery of survey items to assess teachers' pedagogical content knowledge in two areas of the elementary school curriculum—mathematics and writing (Kennedy et al., 1993). As part of this effort, items were written to assess two dimensions of teachers' pedagogical content knowledge: (a) teachers' knowledge of subject matter; and (b) teachers' knowledge of effective teaching practices in a given content area. A report by Deng (1995) discusses the psychometric properties of the various scales constructed as part of this effort. The report shows that TELT researchers' were more successful at measuring teachers' pedagogical content in the area of mathematics than in writing, and more successful in measuring teachers' content

[1] The effect sizes quoted here come from Brewer and Goldhaber (2000: Table 1, page177). The effect size I am using is what Rosenthal (1994) calls an *r*-type effect size. Effect sizes in the *r*-family are designed to express the strength of linear relationships among variables and are suitable for assessing effect sizes in models like linear regression which assume such relationships. Rosenthal's (1994) formula for deriving R^2 from the t-tests in a regression table is the one used here. The formula for deriving *r* (the correlation among two variables) from a t-test statistic is: $r = \sqrt{(t^2/(t^2\text{-df})}$. I simply square this to estimate R^2.

[2] Information on this study—known as the Study of Instructional Improvement—can be found at the following URL: www.sii.soe.umich.edu.

knowledge than knowledge of pedagogy. Overall, however, many of the scales reported in Deng (1995) had low to medium reliabilities and ambiguous face validity. As a result, the efforts of the Michigan State team were encouraging, but for a variety of conceptual and methodological reasons, their work provided little concrete direction for our work.

Conceptual Framework

Lacking firm guidance from previous research, we were forced to work largely from scratch, both in writing survey items to measure teachers' pedagogical content knowledge and in exploring how these items could be used to form scales that would measure important facets of this larger construct. The following decisions were central to the earliest stages of our work and structured the analyses of data reported below:

- Our measures of teachers' pedagogical content knowledge were limited to two, large areas of the elementary school curriculum—reading/language arts and mathematics. Within each of these large curricular domains, however, we attempted to assess teachers' knowledge in relation to a number of more "fine-grained" curricular topics. In reading/language arts, for example, the focal topics chosen for study were word analysis, reading comprehension, and writing. In mathematics, the focal topics chosen for study were number concepts, place value, and operations (with special attention given to multi-digit computation).

- Within each of the "fine-grained" curricular domains just mentioned, we initially identified three dimensions of pedagogical content knowledge to be measured: content knowledge, knowledge of students' thinking, and knowledge of pedagogical strategies. However, while items were written to assess each of these dimensions in the early stages of our work, the analyses reported in this paper focus only on two dimensions: content knowledge and knowledge of students' thinking. Content knowledge is defined here as knowledge of the central concepts, principles, and relationships in a curricular domain, as well knowledge of alternative ways these can be represented in instructional situations. Knowledge of students' thinking is defined here as knowledge of likely conceptions, misconceptions, and difficulties that students at various grade levels encounter when learning various fine-grained curricular topics.

- In all cases, the questionnaire items we developed presented respondents with short—but realistic—scenarios of classroom situations and then posed one or more multiple choice questions about these scenarios, where each multiple choice question contained a "correct" choice and several "incorrect" choices. Decisions about correct and incorrect responses to particular questions were based on research on teaching and learning in the "fine-grained" curricular domains. Figures 1 and 2 provide illustrations of the types of survey items we constructed.

Overall, it should be noted that we view our initial efforts to measure teachers' pedagogical content knowledge as quite limited. For example, while the data reported here focus on two of the most central curricular domains in the elementary school curriculum—reading and mathematics—no attempt was made to enumerate *all* of the fine-grained curricular topics within these larger domains. As a result, we are in no position to claim that.

Figure 1:
An Item Used to Measure Teachers' Knowledge of Student Thinking in the Domain of Reading Comprehension

Two students, Linda and Sean, are given an illustrated children's book to read called *My Pal Mike*. A passage from the book, and Linda and Sean's attempts to read the passage, are presented below. In the book, the passage is accompanied by an illustration, but the illustration is not included here due to space limitations.

A passage from *My Pal Mike:*	**Linda, reading the passage out loud:**
Mike has a big kite.	"Mike had a kite."
Mike runs with his kite.	"Mike runs with the kite."
I like to see the kite rise up and up.	"Mike likes to watch the kite (pause) up (pause) I like to see the kite (pause) fly up."

To what extent do you agree with the following claims about Linda's reading of this passage? ***(Mark one box with an "X" for each item below.)***

	Agree Strongly 1	**Agree Somewhat 2**	**Disagree Somewhat 3**	**Disagree Strongly 4**
7a. Linda is proficient at reading words with consonant blend……	☐	☐	☐	☐
7b. Linda has mastery of a group of common sight words……	☐	☐	☐	☐
7c. Linda uses her knowledge of syntax to help her read text……	☐	☐	☐	☐
7d. Linda monitors for meaning……	☐	☐	☐	☐
7e. Linda gets the important content words correct……..	☐	☐	☐	☐
7f. Linda exchanges visually-similar sight words……..	☐	☐	☐	☐
7g. Linda relies too heavily on phonetic details of text……	☐	☐	☐	☐

Figure 2:
An Item Used to Measure Teachers' Content Knowledge in the Area of Place Value

During a district mathematics workshop, one of the course leaders, Mr. Linden, gave the participating teachers a particularly challenging problem:

> Thinking about tens and ones, 23 is usually written as 2 tens and 3 ones. But it can also be rewritten as 23 ones, or as 1 ten and 13 ones. How many ways can 72 be written?

During a break a few teachers were comparing their attempts to solve the problem. Listed below are several different answers that teachers came up with. Which do you think is correct? (Stem and item, combined)

- [1] 6
- [2] 8
- [3] 7
- [4] 3
- [8] I'm not sure.

Our measurement efforts adequately sample particular domains of the school curriculum within which pedagogical content knowledge is formed and operates. Instead, our approach simply begins with an attempt to measure different dimensions of teachers' pedagogical content knowledge within a limited set of "fine-grained" curricular domains, saving questions about how to sample from (and generalize to) larger curricular domains for a later date.

A similar point can be made about our efforts to measure particular dimensions of pedagogical content knowledge. In theory, at least, we identified three dimensions of this larger construct—teachers' content knowledge, teachers' knowledge of students' thinking, and teachers' knowledge of alternative pedagogical strategies. As discussed earlier, however, this paper discusses scales measuring only two of these dimensions. Moreover, we recognize that other theorists might want to identify additional dimensions of pedagogical content knowledge for measurement. As a result, we make no claims to having adequately sampled all of the possible facets of pedagogical content knowledge that theorists (including ourselves) might identify.

Finally, the analyses presented here are limited by the relatively small size of the sample used to pilot our survey items. Lacking sufficient sample size in the study reported here, we were unable to study in any detailed way the "dimensionality" of our measures of pedagogical content knowledge. As a result, many fundamental questions about the structure of teachers' pedagogical content knowledge remain to be addressed. In the current study, for example, we were unable to investigate in any detailed or satisfactory way whether teachers with superior content knowledge also had superior knowledge of students' thinking, or whether teachers' knowledge in both these areas was consistent across the various "fine-grained" curriculum areas where we developed scales. A detailed analysis of these questions requires analysis of the extent to which our measures form a single measurement dimension or whether they instead form multiple dimensions, but in the present study, we simply lacked sufficient sample size to conduct such analyses. As a result, a better understanding of this problem awaits further research.

Sample and Data

Data for the study reported here were gathered in the summer and fall of 1999. As part of a pilot study designed to develop and validate questionnaire items assessing elementary school teachers' pedagogical content knowledge, we sent 123 elementary school teachers in Michigan and Texas self-administered questionnaires containing items designed to measure teachers' professional knowledge in several "fine-grained" areas of the school curriculum. After three mailings, 104 of these teachers returned completed questionnaires, for a unit response rate of 84.5%. As a group, teachers participating in the study were roughly evenly divided in terms of grade levels taught. All of the teachers held elementary teaching certifications, about half held advanced degrees in education or another field, and about half had been teaching for 15 years or more.

As part of the pilot study, two "alternate forms' of the questionnaire were administered to teachers in the sample. Both forms contained sections designed to measure teachers' pedagogical content knowledge in the larger domains of reading and mathematics. The reading sections of the two forms were roughly parallel in terms of the facets of

pedagogical content knowledge being measured (where these facets were defined as content knowledge and knowledge of pedagogy) and in terms of "fine grained" curricular topics represented. This was less true of the sections on mathematics, however. In the mathematics sections, one questionnaire form was designed primarily for primary grade teachers (K-2) and included "fine-grained" topics taught at this level of the school curriculum, while another form was designed primarily for intermediate grade teachers (3-4) and focused on a different set of "fine-grained" topics.

Most teachers in the study completed only one form of the questionnaire, which was randomly assigned to respondents. However, in order to obtain more robust scaling results, a sub-sample of teachers was asked to complete both questionnaire forms. Overall, 38 teachers completed Form A of the reading questionnaire, 29 teachers completed Form B of the reading questionnaire, and 29 teachers completed both forms. In mathematics, 33 teachers completed Form A (the lower grades questionnaire), 24 teachers completed Form B (the intermediate grades questionnaire), and 26 teachers completed both Forms A and B.

Measurement Strategy

The basic data for the study consist of teachers' responses to multiple choice questions embedded within short (but realistic) classroom scenarios focused on a particular "fine-grained" area of the school curriculum. As discussed earlier, the questionnaire was constructed so that each classroom scenario was followed by one or more multiple choice questions. In all scenarios, any given multiple choice question was designed to measure a single facet of teachers' pedagogical content knowledge, that is, teachers' content knowledge or knowledge of students' thinking. Readers interested in the complete pool of scenarios and items used in this study can consult the original questionnaire forms, which are attached as Appendices C and D of this report.

Item Pool

Tables 1 and 2 (next page) show the kinds of scenarios and items we developed.[3] In an ideal world, we would have scenarios and associated questionnaire items available to measure *both* forms of pedagogical content knowledge for *each* of the fine-grained curricular topics under study. In point of fact, however, we were unable to implement this approach in the pilot study. As Table 1 shows, we did manage to develop scenarios and item pools that measured both types of pedagogical content knowledge across all of the fine-grained curricular topics in mathematics. However, as Table 2 shows, the types of pedagogical content knowledge being measured in reading/language arts are distributed unevenly across fine-grained curricular topics.

[3] In this paper, we define an "item" as any response option embedded within a multiple choice question that was scored as "correct" or "incorrect" for measurement purposes. Using this approach, for example, Figure 1 above includes one scenario, a single multiple choice question, and seven items; Figure 2 above includes one scenario, a single multiple choice question, and one item.

Table 1: Number of Items Assessing Teachers' Pedagogical Content Knowledge in Mathematics

	Facet of Pedagogical Content Knowledge	
	Content Knowledge	Knowledge of Students' Thinking
• Number Concepts	10 items (3 scenarios)	4 items (1 scenario)
• Place Value	4 items (4 scenarios)	19 items (4 scenarios)
• Operations	1 item (1 scenario)	4 items (1 scenario)
• Multi-digit Computation	17 items (5 scenarios)	12 items (3 scenarios)

Table 2: Number of Items Assessing Teachers' Pedagogical Content Knowledge in Reading/Language Arts

	Facet of Pedagogical Content Knowledge	
	Content Knowledge	Knowledge of Students' Thinking
Word Analysis		
• Letter-sound relationships	13 items (3 scenarios)	
• Phonemes	8 items (1 scenario)	
• Word recognition/sight words		12 items (4 scenarios)
• Phonetic cues		6 items (4 scenarios)
• Context/Picture/Syntactical Cues		16 items (4 scenarios)
Reading Comprehension		
• Monitoring for meaning		4 items (4 scenarios)
Writing		
• Editing process	5 items (2 scenarios)	

Measurement Model and Approach to Building Scales

Our approach to building scales from the item pools described in Tables 1 and 2 involved using the computing program BILOG, a commercially-available program for item analysis and scoring. Using this program, we fitted items to scales using a Rasch measurement model. The Rasch model is one of a family of measurement models described in item-response theory (IRT). In essence, Rasch models are "one-parameter" IRT models, that is, models that estimate a single parameter for each item in a scale—a level of difficulty. In the model, all items are weighted equally in determining the ability measure for a particular person, and a person's measured ability on a given scale is a direct (1 to 1) function of the number of items answered correctly.

An important feature of the Rasch measurement model is the strict assumption of unidimensionality of measurement. In this one-parameter IRT model, all of the items in a given scale are assumed to have the same item-to-trait correlation, implying that all items are measuring the same underlying construct in exactly the same way (other IRT models relax this assumption by including a discrimination parameter and/or a guessing parameter). While it would be desirable to build scales using more complex IRT models, Rasch models can be fit with as few as 25 subjects, whereas more complex, two and three-parameter IRT models require as many as 500 to 1000 subjects. Lacking sufficient sample size to fit more complex IRT models, we were restricted in the present study to using Rasch models to build scales.

Scales were built in a series of stages. Given the strict assumption of unidimensionality, we began our measurement efforts by building scales at the most discrete level of measurement, where a single facet of teachers' pedagogical content knowledge (i.e., teachers' content knowledge or teachers' knowledge of students' thinking) is measured in reference to a single "fine-grained" area of the school curriculum. This corresponds to building scales in each of the cells in Tables 1 and 2 where items are available. Once initial scales were formed using all of the items in a cell, we examined the item-to-scale correlations in an attempt to identify items that did not "fit" the Rasch model. In this stage of the analysis, the item with the lowest item-to-scale biserial correlation was omitted from a scale and a Rasch model was fit to the remaining items. This procedure was repeated until the reliability of the scale failed to improve. Once these "fine-grained" and facet-specific scales were built, we moved to building scales at a broader level of analysis. Here, we combined items from scales measuring teachers' knowledge at a more fine-grained level into more encompassing measures of teachers' pedagogical content knowledge—for example, scales that combined facets of teachers' pedagogical content knowledge into a single scale within a particular fine-grained curriculum area, or scales that focused on a single facet of knowledge but that were aggregated across our "large" curriculum domains.

In all analyses, our main goal was to build scales with the highest levels of reliability possible. In IRT models, reliability is defined as:

$$\rho = \frac{\text{var}(\theta)}{\text{var}(\theta) + \text{averege}(\text{se}(\theta))^2}\ , \qquad (1)$$

where ρ is the test reliability, θ is the scale score for a person, se (θ) is the estimated standard error of the scale score for a person, and average (se(θ)) is the se(θ) averaged over the distribution of θ. Note that in a Rasch model, reliability increases when items with low item-to-scale correlations are deleted because weighting such items equally with items having higher item-to-scale correlations effectively increases (se(θ)). Also note that if we were to fit a more complex, two-parameter IRT model to our data, deleting items with low item-to-scale correlations might increase the reliability above that obtained by deleting the items in the one-parameter Rasch model. In this sense, then, the reliabilities that we report for scales in this paper can be interpreted as an approximate lower bound to the reliability that could be obtained were we to use a more complex, two or three-parameter IRT models to fit the data to scales.

Overview of Results

Detailed results of our scaling work are presented in Appendices A and B (attached to this report). These appendices describe the content of each of the items included in our item pools and in the final scales, as well as a map showing where these items can be found in the pilot study questionnaires, which are also attached to this report. In addition to providing detailed information on the items used in our work, the appendices also show: (a) which of the items from the initial pool of items (shown in Tables 1 and 2) were kept and which were deleted in the forming of particular scales; (b) the item-to-scale biserial correlations for "kept" and "deleted" items; and (c) the overall scale reliabilities for each of the scales that we constructed. Readers interested in arriving at a full understanding of our work are strongly urged to examine these appendices carefully.

For purposes of brief presentation, however, an overall summary of the measurement results is shown in Tables 3 and 4 (below). These tables show the number of items and overall reliabilities for each of the scales that we constructed from the item pools discussed in Tables 1 and 2. The results are summarized separately for scales in mathematics and for scales in reading/language arts.

Mathematics Scales

Table 3 (next page) shows that we constructed eleven different scales in the curricular area of mathematics and that, overall, our scaling results were quite mixed. As the table shows, we were successful in constructing at least some scales with sufficient reliablities in some of the "fine-grained" curricular areas of interest. The table also shows that our successes could be found across both of the facets of pedagogical content knowledge that we were attempting to measure.

The data reported in the first column of Table 3 show the scaling results for measures of teachers' mathematics *content knowledge*. Here, we developed several different scales with widely varying levels of reliability. These included a 4-item scale measuring teachers' content knowledge in the "fine-grained" area of number concepts that had a reliability of 0.674, and a 14-item scale of teachers' content knowledge in the fine-grained area of multi-digit computation that had a reliability of 0.859. However, the 4-item scale we

Table 3. Scales and Reliabilities for Teachers' Professional Knowledge in Mathematics

	Content Knowledge	Knowledge of Students' Thinking	Pedagogical Content Knowledge
Number Concepts	0.674 (4 items)	0.522 (3 items)	0.500 (8 items)
Place Value	0.000 (4 items)	0.764 (13 items)	0.767 (13 items)*
Operations		0.545 (4 items)	
Multi-Digit Computation	0.859 (14 items)	.0744 (5 items)	0.874 (20 items)
Overall Content Knowledge	0.869 (23 items)		
Overall Knowledge of Students' Thinking		0.785 (24 items)	
* No content knowledge items were kept in this scale.			

constructed to measure teachers' content knowledge in the curricular area of place value had a reliability of 0.00(!). Overall, Table 3 also shows that we constructed a 23-item scale to measure teachers' content knowledge in the area of mathematics generally and that this scale had an acceptable reliability of 0.869.

The second column of Table 3 shows the results for measures of the facet we called teachers' *knowledge of students' thinking* in mathematics. Here too, we experienced varying success in measurement construction. A 5-item scale measuring teachers' knowledge of students' thinking in the area of multi-digit computation had a reliability of 0.744, and a 13-item scale measuring teachers' knowledge of students' thinking in the area of place value had a reliability of 0.764. These are acceptable reliabilities, especially if one accepts the argument that our one-parameter Rasch model provides us with "lower bound" estimates of reliability. However, not all of the scales measuring teachers' knowledge of students' thinking had these levels of reliabilty. For example, a 4-item scale measuring teachers' knowledge of students' thinking about operations, and a 3-item scale assessing teachers' knowledge of students' thinking about number concepts, had reliabilities in the range of 0.50. When all of the "student thinking" items were combined across the different "fine-grained" curricular areas under consideration, we developed an overall scale measuring teachers' knowledge of students' thinking in mathematics that included 24 items and that had a reliability of 0.785.

The third column of Table 3 displays the results of our attempts to combine the different facets of teachers' content knowledge and teachers' knowledge of student thinking into overall measures of teachers' *pedagogical content knowledge* in mathematics. Three such scales were constructed in the fine-grained curricular areas of number concepts, place value, and multi-digit computation. The results show that we had only modest success in this endeavor. For example, the place value scale shown in the third column shows that the overall scale intended to measure teachers' pedagogical content knowledge in place value included only "content knowledge" items—with no "student thinking" items included in the final scale. Therefore, the combined measure is virtually the same measure as presented in the knowledge of students' thinking column. Also, an examination of the reliability for the 8-item number concepts scale shows that this overall measure of teachers' pedagogical content knowledge had a lower reliability (0.500) than the reliabilities for its two component

scales (content knowledge and knowledge of students' thinking), each of which also had fewer items than the combined scale. Finally, the 23-item scale measuring teachers' pedagogical content knowledge in the area of multi-digit computation had a reliability of 0.874, exceeding the reliabilities of its two component subscales. However, the Spearman-Brown prophecy formula predicts an increase in reliability to at least 0.900 given the addition of six items to either of these subscales. Overall, then, the results presented here begin to suggest that (within the fine-grained curricular areas under consideration here), the two separate facets of teachers' pedagogical content knowledge that we have attempted to measure might not be as strongly related as anticipated, raising questions about the unidimensionality of any overall measure of "pedagogical content knowledge" that might be constructed.

Reading/Language Arts Scales

The results of our measurement work in the curriculum area of Reading/Language Arts are presented in Table 4 (next page). As discussed earlier, in this curriculum domain we did not develop items to measure both facets of teachers' pedagogical content knowledge in each of the fine-grained curricular areas under consideration. Instead, we developed measures of teachers' content knowledge in one set of fine-grained curricular areas and measures of teachers' knowledge of students' thinking in a different set of fine-grained curricular areas. This approach is less than ideal—limiting our ability, for example, to construct "overall" scales of pedagogical content knowledge in the curricular domains under study—but resulted from the challenge of developing sufficient numbers of items in a curricular area where previous conceptual and measurement work on the nature of teachers' pedagogical content knowledge is largely absent. Moreover, our work in this area concentrated on assessing teachers' content knowledge or knowledge of students' thinking in the larger area of the Reading/Language Arts curriculum often called "word analysis" and/or "phonics." For example, six of the eight curriculum domains in which measures were developed were in this area.

The first column of Table 4 presents the results for our measures of teachers' *content knowledge* in Reading/Language Arts. In general, these measures have acceptable (and sometimes very high) levels of reliability. For example, the 8-item scale measuring teachers' content knowledge of letter/sound relationships had a reliability of 0.697, the 7-item scale measuring teachers' content knowledge of phonemes had a reliability of 0.999, and a 12-item "word attack" scale combining items from these two areas had a reliability of 0.911. Only the 5-item scale measuring teachers' content knowledge in the area of editing (a topic in the writing curriculum) had an *un*acceptable level of reliability (0.109). We also tried to develop an "overall" measure of teachers' content knowledge in Reading/Language Arts by combining items from the editing scale with items from the scales measuring teachers' content knowledge in letter/sound relationships and phonemes. Clearly, this scale does not sample all of the fine-grained curricular domains in the field of Reading/Language Arts very completely. Moreover, although this 21-item, summary scale had an acceptable reliability of 0.870, this reliability is lower than the reliability of the combined "word attack" scale, bringing into question the idea that the "overall" scale is measuring a single dimension of teachers' content knowledge scale in Reading/Language Arts.

Table 4. Scale Reliabilities for Teachers' Professional Knowledge in Reading

	Content Knowledge	Knowledge of Students' Thinking	Pedagogical Content Knowledge
Letter/Sound Relationships	0.697 (8 items)		
Phonemes	0.999 (7 items)		
Word Attack	0.911 (12 items)		
Editing	0.106 (5 items)		
Overall Content Knowledge	0.870 (21 items)		
Word Recognition/Sight Words		0.486 (6 items)	
Use of Phonetic Cues		0.374 (3 items)	
Use of Context, Picture, and Syntactical Cues		0.724 (11 items)	
Monitors for Meaning		0.433 (4 items)	
Overall Knowledge of Students' Thinking		0.798 (28 items)	

The second column of Table 4 shows the measurement results for the scales we developed measuring teachers' *knowledge of students' thinking* in the curricular area of Reading/Language Arts. Once again, the scales focus largely on areas of reading arts commonly known as "word analysis" and/or "phonics", although we also developed one scale in the area of reading comprehension (monitors for meaning). The results are mixed, as Table 4 shows. Only one of the scales measuring teachers' knowledge of students thinking in the areas of "word analysis/phonics" obtained an acceptable level of reliability—the 11-item scale measuring teachers' knowledge of students' thinking in the area of using context, picture, and syntactical cues (reliability=0.724). By contrast, the 6-item word recognition/sight words scale had a reliability of only 0.486, and the 3-item use of phonetic cues scale had a reliability of only 0.374. Similarly, the one scale we constructed in the area of teachers' knowledge of students' thinking in the area of reading comprehension—the 4-item, monitors for meaning scale—had a reliability of 0.433. Still, when we combined items across all four of the fine grained curriculum areas into an overall scale measuring teachers' knowledge of students' thinking in the area of reading language arts, the resulting 28-item scale had a reliability of .798, a reliability which no doubt could be improved by use of a two- or three-parameter IRT measurement model rather than the one parameter, Rasch model used here.

Discussion

The results just presented show that our success in using survey items to construct reliable scales measuring teachers' pedagogical content knowledge was decidedly mixed. On one hand, we did succeed in developing a number of scales (containing as few as 6 – 10 items) that reliably measured particular facets of teachers' pedagogical content knowledge in particular "fine-grained" areas of the elementary school reading/language arts and mathematics curricula. But these successes must be balanced against our inability to develop reliable scales in other curricular domains. Overall, for example, we built 22 scales. Two of these had reliabilities above 0.90, another 3 had reliabilities above 0.80, another 7 had reliabilities above 0.70, and 10 had reliabilities below 0.70. In one sense, our results can be seen as a kind of "existence proof" demonstrating that it *is* possible to develop reliable scales measuring particular facets of teachers' pedagogical content knowledge in fine-grained areas of the school curriculum. But the results also serve as a cautionary tale about the difficulties involved in this effort.

One difficulty we faced was developing items (and scenarios) that adequately tapped the full range of underlying "abilities" or "levels" of teachers' content and pedagogical knowledge in the various curricular domains under study. For example, a fairly large proportion of the items that we developed were either "too easy" or "too hard" for the vast majority of the teachers in our sample, meaning that the vast majority of teachers in the sample either answered these items correctly or answered these items incorrectly (see, for example, the p-values for specific items reported in Appendices A and B). This has important consequences for scale construction. As an example, three of the five items in our scale measuring teachers' knowledge of students' thinking in the area multi-digit computation had p-values greater than .875. The result is the construction of scale with reasonable reliability, but one that is also discriminating effectively only between poorly performing teachers and the rest of the teacher pool.

Yet another problem we confronted was in writing items and scenarios that provided clear and sufficient information to respondents. An inspection of the survey instruments attached to this report, as well as the item-to-scale biserial correlations presented in the appendices, suggests that a number of scenarios and/or items that we developed suffered from ambiguity or presented information about classroom situations that was simply too "thin" for respondents to make informed responses. Thus, an inspection of the detailed data presented in Appendices A and B will show that a great many of the items that were developed had very low (and occasionally negative) item-to-scale biserial correlations. One consequence was that, in building scales, we were forced to drop many of the items embedded in a particular scenarios, making the use of such scenarios in multi-purpose surveys (where space is at a premium) inefficient. Clearly, future item- and scenario-writing efforts need to overcome these problems, and as a result, we have developed a guide to successful item-writing that is attached to this report as Appendix C.

Another set of problems that we encountered stem from the small sample of teachers achieved in this study. A verification (and extension) of our results with a larger sample of teachers would be helpful for several reasons. One is that we sometimes were forced to delete particular items from our scales because of the aberrant responses of just a few teachers. This was particularly true of items with extremely high or extremely low p-values,

since under these conditions, it takes only one or two teachers with unexpected responses to produce low item-scale biserial correlations for that item. Clearly decisions to delete items should be based on a larger pool of teachers.

Equally important, a larger sample of teachers would have allowed us to move beyond the use of a simple, one-dimensional Rasch model and to employ more realistic IRT models that reflect the reality of the constructs we are attempting to measure. In light of the fact that we are attempting to develop measures of two facets of teachers' pedagogical content knowledge, and to do so in a variety of "fine-grained" curricular domains, it is probably unrealistic to assume that that a simple, one-parameter, Rasch model can be directly applied to all of the constructs we are attempting to measure. In fact, the relatively severe assumptions associated with the Rasch model might account (at least in part) for the large number of items we found it necessary to delete during our scale construction work, both in the development of measures of a single facet of pedagogical content knowledge in fine-grained curricular domains, but also in our single and multi-faceted measures where we aggregated data across larger curricular domains. In our current work, we are now using much larger samples of teachers in a continuing effort to pilot items and develop scales. At a minimum, we will be using such data to explore the use of two- and three-parameter IRT models, allowing items with lower biserial correlations to be included in our scales without reducing the reliability of the resulting measures.

More importantly, however, the larger sample size we are using in our research will allow us also to explore in a much more detailed and satisfactory way the "dimensionality" of the constructs we are attempting to measure. In this paper, we adduced some evidence of a lack of unidimensionality in scales—for example, measures combining different facets of teachers' professional knowledge (e.g., content knowledge and knowledge of students' thinking) across curricular areas, or measures combining a single facet of teachers' knowledge across curricular domains, often had lower reliabilities than did the scales constructed at lower levels or lower reliabilities than would be expected from the Spearman-Brown formula. All of this suggests that there may be distinctive (and only modestly correlated) a dimension of teachers' pedagogical content knowledge, but more sophisticated scaling work is needed to confirm this initial evidence.

Clearly, an examination of the dimensionality of the constructs being measured should be an important aspect of future work on the measurement of teachers' pedagogical content knowledge. This is critical, not only for our understanding of the constructs themselves, but also because the standard errors of measurement and the reliabilities upon which the scales are based in this paper rely on the assumption of unidimensionality. If that assumption is violated, the use of a Rasch model with multidimensional items tends to overestimate standard errors and underestimate the variability of the construct in question, resulting in low reliabilities. Thus, clearly determining the dimensionality of the scales will be critical in future work, not only to better understand the "structure" of teachers' pedagogical content knowledge, but also to measure it well in future survey research efforts.

Conclusion

We began this paper with a discussion of the need for survey researchers in the field of education to move way from their current reliance on measures of general cognitive

ability or proxy measures of teachers' professional knowledge in research on teaching in order to develop measures of teachers' professional knowledge that are more in line with conceptions of teachers' pedagogical content knowledge now emerging in current research on teaching and in high-quality programs of teacher assessment. Our efforts in this regard—limited as they are—can be seen as one effort to respond to this call.

In one sense, the results of our initial pilot study into the measurement of teachers' pedagogical content knowledge are encouraging. We have shown, for example, that it is possible to reliably measure particular facets of teachers' pedagogical content knowledge in very fine-grained areas of the school curriculum with as few as 6-10 survey items. However, our pilot study also suggests that future efforts to develop survey-based measures in this area will be challenging. For one, the limited number of scales that we reported on in this paper do not begin to scratch the surface of the many curricular areas where teachers' pedagogical content knowledge is formed and operates. Moreover, if our experience is any indication, developing survey items to measure teachers' pedagogical content knowledge in these areas will require the sustained effort of many researchers. As we discussed here, our own efforts resulted in many survey items being discarded—either because they failed to adequately discriminate the "levels" of knowledge possessed by teachers, or because they failed to provide respondents with the kind of clarity needed for respondents to unambiguously and reliably demonstrate their knowledge. Our efforts therefore suggest that a great deal more effort in developing item banks will be needed before survey researchers can measure teachers' pedagogical content knowledge well.

But even with such item banks, a great deal remains to be learned about the actual nature of teachers' pedagogical content knowledge. In particular, our own conception of this construct suggests that it is multifaceted in nature—involving both knowledge of content and knowledge of students' thinking, facets that develop within a large number of discrete curricular domains. Whether we can ever arrive at summary measures of teachers' pedagogical content knowledge—and whether such summary measures can be developed with a relatively small number of survey items—remains an open question that we (and we hope others) will be exploring in the future. But only more work on basic measurement problems will address these questions, and they are questions that deserve being addressed. In our own work, for example, we want to know if teachers' with strong knowledge of how to represent content to students necessarily also have a strong knowledge of students' thinking, and whether teachers who have strong knowledge of these sorts in one curriculum area also have it in other curriculum areas. We suspect that answers to these questions are related to how such knowledge develops over time, both as a result of teachers' classroom experience and as a result of deliberate steps taken by various kinds of educational organizations to enhance such knowledge. But detailed investigation of these issues appears to require more sophisticated approaches to measurement.

There is, then, an ambitious research agenda for survey researchers interested in studying teachers' pedagogical content knowledge. Full realization of this agenda, however, requires better measures of the central construct. This paper is offered as a first step along the road to developing such measures, but much more work is required before survey researchers can develop and use such measures in future research.

References

Brewer, D.J. & Goldhaber, D.D. (2000). Improving longitudinal data on student achievement: Some lessons from recent research using NELS:88. In, D.W. Grissmer & J.M. Ross (Eds.), Analytic issues in the assessment of student achievement. Washington, DC: U.S. Department of Education.

Coleman, J.S. et al. (1966). Equality of educational opportunity. Washington, D.C.: U.S. Government Printing Office.

Darling-Hammond, L., Wise, A.E., & Klein, S. P. (1995). A license to teach: Building a profession for 21st-century schools. San Francisco, CA: Westview Press.

Deng, A. (1995). Estimating the reliability of the questionnaire used in the Teacher Education and Learning to Teach study. Technical Series Report 95-1. East Lansing, MI: National Center for Research on Teacher Education.

Ferguson, R.F. & Brown, J. (2000). Certification test scores, teacher quality, and student achievement. In D.W. Grissmer & J.M. Ross (Eds.) Analytic issues in the assessment of student achievement. Washington, DC: U.S. Department of Education.

Greenwald, R., Hedges, L.V., & Laine, R.D. (1996). The effect of school resources on student achievement. Review of Educational Research, 66, 361-396.

Kennedy, M. et al. (1993). A guide to the measures used in the Teacher Education and Learning to Teach study. East Lansing, MI: National Center for Reserch on Teacher Education.

Monk, D. H. (1994) Subject area preparation of secondary mathematics and science teachers and student achievement. Economics of Education Review, 13(2), 125-45.

Porter, A.C., Youngs, P. & A. Odden. (In press). Advances in teacher assessments and their use. In, V. Richardson (Ed.), Handbook of Research on Teaching (4th ed.).

Rosenthal, R. (1994) Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis. New York: Russell Sage Foundation.

Rowan, B., F.S. Chiang, and R.J. Miller. (1997). Using research on employees' performance to study the effects of teachers on students' achievement. Sociology of Education, 70, pp. 256-284.

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), 1-22.

Smith, M. & George, L. (1992). Selection methods. In, C.L. Cooper & L.T. Robinson (Eds.), International Review of Industrial and Organizational Psychology, 7, 55-97.