AMERICAN INSTITUTES FOR RESEARCH

MEASURING TEACHER KNOWLEDGE OF THE NRP: AN INSTRUMENT AND PILOT TEST RESULTS

December 30, 2005

Prepared by:
American Institutes for Research
1000 Thomas Jefferson St. NW
Washington, DC 20007-3835

Prepared for:
U.S. Department of Education
Tracy Rimdzius
555 New Jersey Ave., NW, Room 500K
Washington, DC 20208-5500

Table of Contents

Introduction

This report describes a pilot test of a set of items designed to assess pre-service teacher knowledge of the five critical components of early reading instruction as defined by the National Reading Panel (NRP). This report is intended to supplement the *Revised Study Design* document, which describes the development and selection of these items. Therefore, this report will not review the item development process. It focuses specifically on the properties of the items we propose to use for assessing pre-service teacher knowledge.

Although item development will not be revisited here, it is important to note that the items we propose to use (108 multiple-choice and 24 constructed-response items) were part of a larger pool of items that were pilot tested. Table 1 lists the total number of items pilot tested, the different item types, and the components that each item was developed to measure. The four components listed in Table 1 comprise the teacher knowledge of student content engagement (TK-SCE) framework, which was described

Table 1. Pilot Test Items, by Type and Component

	Component 1: Subject Matter Content Level	Component 2: Occasion for Processing	Component 3: Physiological Readiness	Component 4: Motivation	Cross-Component	*Total*
Likert Scale	28	23	15	62	–	*128*
Multiple-Choice	72	48	21	31	–	*172*
Constructed-Response	25	15	5	9	–	*54*
Situational Judgment	15	24	14	22	–	*75*
Q-Sort / Checklist	10	8	–	–	–	*18*
High Fidelity Simulation	–	2	2	–	2	*6*

in the *Revised Study Plan*. The 108 multiple-choice and 24 constructed-response items that are proposed for the teacher knowledge assessment were aligned with the five NRP components during item development and are viewed as subcomponents of the Subject Matter Content Level and Occasions for Processing constructs (refer to Table 2). Q-sort and Likert items were deemed as inappropriate for the pre-service teacher assessment because these items are not objectively scored, while the hi-fidelity and situational judgment items are too experimental in nature to be considered at this point.

Table 2. Number of Items Aligned with the NRP Components

NRP Component	Multiple Choice	Constructed-response	TOTAL
Phonemic Awareness	16	3	19
Phonics	22	4	26
Fluency	10	5	15
Vocabulary	33	8	41
Comprehension	27	4	31
TOTAL	108	24	132

The main objective of the pilot test, therefore, was to collect preliminary data on the items presented in Table 1. For this report, we will only examine data on the items presented in Table 2. The goal of these analyses will be to answer the following questions:

- Are the items appropriately difficult?
- Are the multiple-choice distractors functioning correctly?
- Are the items reliable?
- Does item difficulty vary by level of experience?

The answers to these questions will inform the technical working group (TWG) and Department of Education (ED) regarding the properties of the proposed assessment so that an informed decision can be made about exercising the optional tasks associated with this project.

Method

Participants

Participants in the pilot test were selected to be reflective of teachers who might take an operational version of the TK-SCE survey, not a teacher knowledge assessment to be administered to pre-service teachers. Therefore, all participants except for two had experience teaching in the primary grades. Basic demographics on these participants are described below.

Teacher Recruitment

Current or recent primary grade teachers were recruited to participate in the pilot test. All teachers who had taught kindergarten, 1st, or 2nd grade in a public school in the last three years were eligible. The only requirement was that teachers were willing and available to participate in the pilot test for the entire four-hour period.

Five locations were chosen for pilot testing in order to have representation from different parts of the country. Teachers from each area were eligible for participation. The five areas were:

- Raleigh / Durham, North Carolina
- Chicago, Illinois
- St. Louis, Missouri
- Dallas, Texas
- San Diego, California

A letter, fact sheet, and description of the study were sent to all district superintendents in the selected sites. About a week after the mail out, telephone interviewers began contacting superintendents to recruit and schedule districts. Recruitment was slow going at first. Interviewers faxed and re-mailed the materials as requested and made numerous call backs before finally getting answers about participating in the pilot study. Some districts were very eager to participate and saw the teacher incentive as a great opportunity for their teachers to earn a little extra money. Other districts were not interested and were not motivated by the teacher incentive.

As sessions were scheduled the date, time, location, and contact person's information were entered into the receipt control system. Teachers' names, emails, and phone numbers were also recorded so that checks could be requested in advance of the sessions. This information was provided to the interview teams in advance of the sessions so that they could send reminder emails to teachers.

Teacher Demographics

A total of 589 teachers participated in the pilot test. Participating teachers were distributed across each of the five geographic regions as reflected in Table 3.

Table 3. Number of Teachers per Region

Region	State	Number of Sessions	Number of Teachers
San Diego	CA	11	50
Dallas	TX	19	173
St. Louis	MO	13	125
Chicago	IL	12	100
Raleigh / Durham	NC	17	141

The vast majority of the participants in the pilot test were female (98%), which is representative of elementary school teachers in the US. In providing demographic data, 10% of participants identified themselves as Black or African American; about 6% identified themselves as Hispanic; and about 81% identified themselves as White. Regarding age, 35% were 26-35, 22% were 36-45, and 25% were 46-55. The vast majority of participants reported having at least a Bachelor's degree, and many stated that they had a Master's degree. Most of the participants majored or minored in Elementary Education. Individuals in the sample also reported having significant teaching experience in early elementary (82% with four or more years) and upper elementary grades (82%

with four or more years). Finally, virtually all of the participants had some sort of teaching certificate, including a few who were working toward (4%) or had attained (3%) their National Board certification.

Test Versions

For the purpose of the pilot test, the pool of items was split and two alternate versions of the survey were created (Version 1 and 2). We had to create separate versions due to the total number of items written and the desire to pilot as many items of this pool as possible.

Although creating alternative versions of the survey allowed us to collect data on as many items as possible, it did create some challenges. For example, because no individual completed every item on Version 1 and 2, items from the two versions could not be correlated with each other. For instance, the 108 multiple-choice and 24 constructed-response items that are the focus of the current report were distributed so that 56 multiple-choice and 12 constructed-response items were on Version 1 and 52 multiple-choice and 12 constructed-response items were on Version 2. Thus, we pilot tested two shorter versions of the teacher knowledge assessment (i.e., alternative forms).

In addition to creating two alternate versions of the survey, we counter-balanced sections of the survey within each version to guard against order and fatigue effects. For example, we did not want any particular item type to always appear last and hence not be reached. This counterbalancing process produced five differently ordered forms of each version of the survey (i.e., 10 unique forms in total). Table 4 presents the number of teachers who received each form.

Table 4. Number of Teachers per Form

Form	Version	N
1	1	62 teachers
2	2	64 teachers
3	1	66 teachers
4	2	77 teachers
5	1	63 teachers
6	2	66 teachers
7	1	56 teachers
8	2	55 teachers
9	1	36 teachers
10	2	44 teachers

Similar to the creation of Version 1 and 2 of the survey, the counterbalancing of items within each version had advantages and disadvantages. On the positive side, we guarded against order and fatigue effects, which are common with long assessments. Also, we were able to collect some data from some respondents on each item that was developed. On the negative side, splitting the item pool in half and counterbalancing produced smaller than desirable numbers of respondents for some items. Sample size per items and our approach to dealing with this issue is discussed in the results section.

Procedure

Two individuals administered the survey in each location; thus, a total of ten administrators were used. All of these administrators had previous experience with various data collection projects.

Survey administrators completed a two-day comprehensive training course on how to conduct the pilot test. During this training session, the project was introduced and the procedures were described in detail. In addition, much of the training involved familiarizing the administrators with the computers and the application that was used for data collection. Administrators spent time practicing the computer set-up process and the data saving procedures. The Administrator Guide that was used in training is available upon request.

Data collection occurred between September and November, 2005. All forms of the survey were administered to participating teachers on laptop computers. Computers were not connected to the Internet or to a network, but operated as independent machines with the software containing the items resident on each laptop. Teacher responses were directly saved to the hard drive on the laptop. Upon completion of the survey, the administrators saved the results on blank CD's, via the CD-ROM drive which was built into all of the computers.

In most cases, data collection occurred at local schools that volunteered to provide meeting space. Up to ten participants were scheduled for each session and given instructions about the project. The two administrators were scheduled to arrive one hour before data collection was to begin. During this time, they introduced themselves to school personnel and set-up the meeting room for data collection. This mainly involved setting up the laptops in the room. Most sessions occurred either after school or on the weekends; because of this and the time requirements, food was provided for participants.

Each pilot test session was four full hours and the four-hour session was broken down into five smaller, time-limited test sections. At the scheduled start time, the administrators commenced the check-in procedures and gave an overview of the project. Then, participants started the survey. The first section for everyone was the Opinion (Likert) items. Upon completion of that, they started Section 2, the content of which varied by the form that each individual was completing. Section 1 and Section 2 took a combined 80 minutes after which there was a ten-minute break. Section 3 lasted 50 minutes, which was followed by a ten-minute break. Section 4 also took 50 minutes, and was directly followed by Section 5, the background section and check-out, which lasted 20 minutes. Because of the large number of items allocated to each section, very few participating teachers were able to answer all of the items of a given form.

Scoring & Analysis Plan

This section describes the analyses that were conducted. Prior to analyzing the data the items needed to be scored. The scoring process for the multiple-choice and constructed-response items is described next.

Item Scoring

Multiple-Choice Items

We designed the multiple-choice items to have one clear, best response. Participants received credit for selecting the right choice out of the alternatives provided (A, B, C, or D). Participants were not instructed that they would be penalized for skipping or failing to complete a certain number of items due to time. As a result, respondents varied significantly in the number of multiple-choice items they actually completed.

Constructed-Response Items

The constructed-response items required the participants to respond in writing to open-ended questions. While this item format measures a unique type of knowledge that is different from that measured by multiple-choice items, it brings with it some clear challenges when scoring the items. The primary challenge involves having raters score the responses in a standardized, reliable, and valid manner. In response to this challenge, we devised an approach that utilizes specific scoring protocols, multiple raters, and expert judges. During the development of the constructed-response items, item writers created scoring rubrics, or standardized scoring keys, that describe how each item should be scored. Following data collection, raters scored the items using these rubrics, which defined correct and incorrect answers. Thus, raters were to make judgments as to whether the response was deemed correct (2 points), partially correct (1 point), or incorrect (0 points or no credit). Raters consisted of nine judges. Three of these raters were subject matter experts in the field of elementary school teacher education or early reading instruction while six raters were research assistants working on the project.

All raters were trained to use the rubrics and the scoring program. During the training, raters were provided with several items to score and examples of acceptable and unacceptable responses. The raters scored all the items independently and then convened to discuss their scores and the rationale for their decisions. Through discussion, raters began successfully reaching consensus on ratings. The process was repeated and the raters made progress in their observations and rationales. The goal of the training was to improve judgments and accuracy by teaching the raters to share similar schemas of correct and incorrect responses. After being trained, preliminary reliability and accuracy checks were conducted prior to commencing the actual constructed-response scoring. Interclass correlations, percent agreement among raters, correlations among raters, and agreement indices between the six research assistants and the three subject matter experts were calculated.

The results suggest that the raters were reliable and accurate, which increased our confidence in the quality of the item scoring. Table 5 presents the intra-class correlation obtained after rater training. Conventions based on past research suggest that intra-class correlations less than .40 are considered "poor," between .40 and .59 are considered "fair," .60 to .74 are considered "good" and intra-class correlations above .74 are considered "excellent." The results show that most of the obtained intra-class correlations were in the good or excellent categories.

Table 5. Intra Class Correlations among Raters Obtained After to Rater Training

Partners	Intra-class Correlations
Rater 1 and Rater 2	0.92
Rater 1 and Rater 3	0.90
Rater 4 and Rater 2	0.89
Rater 4 and Rater 3	0.85
Rater 5 and Rater 4	0.74
Rater 5 and Rater 6	0.70
Rater 7 and Rater 1	0.69
Rater 6 and Rater 8	0.68
Rater 6 and Rater 3	0.67
Rater 9 and Rater 2	0.66
Rater 9 and Rater 7	0.65
Rater 5 and Rater 8	0.61
Rater 4 and Rater 7	0.61
Rater 6 and Rater 7	0.61
Rater 6 and Rater 2	0.65
Rater 1 and Rater 8	0.56
Rater 9 and Rater 3	0.56
Rater 5 and Rater 3	0.53
Rater 5 and Rater 9	0.53
Rater 5 and Rater 2	0.49
Rater 5 and Rater 1	0.44
Rater 4 and Rater 8	0.25
Average ICC:	0.64

Table 6 presents the percent agreement among raters after training. These data show that raters agreed more than 83% of the time. These findings further demonstrate the effectiveness of rater training and that the constructed-response items can be reliably scored.

Even though the raters were found to be reliable, all of the constructed-response items were scored by at least two raters as a final consistency check. Scores for each participant were calculated by computing the average score of the two ratings. Scores for each item ranged from 0 to 2.

Table 6. Average Percent Agreement

Item ID:	Average Percent Agreement
sam_01	79.6
sao_14	82.4
sao_17	87.4
sap_01	86.4
sas_04	76.6
sas_05	82.6
sas_08	92.6
sas_09	87.2
sas_11	82.5
sas_15	73.1
sas_16	88.0
All Items	83.5

Analysis Procedures

Based on the goals of this pilot test, we analyzed all of the items for difficulty and discrimination as well as analyzed the extent to which the items possessed internal consistency. However, the multiple-choice and constructed-response items required somewhat different approaches for meeting these goals. Below we describe our approach to analysis for each item type. This section is followed by the results of the pilot test.

Multiple-Choice Items

Descriptive statistics and reliability analyses were conducted on the multiple-choice items. Regarding descriptive statistics, the percentage of respondents who answered each multiple-choice item correctly was calculated (i.e., item difficulty) and the number of respondents who selected each response option was determined to assess the quality of the distractors. Regarding reliability, alpha was calculated for each set of multiple-choice items (alpha for the items that appeared on Version 1 and an alpha for the items that appeared on Version 2) that we propose to measure pre-service teacher knowledge of the NRP. We also estimated alpha if all the multiple-choice items were included on a single version of the assessment using Spearman-Brown. Finally, item-total correlations were calculated as part of the reliability analysis (an indicator of item discrimination) and the impact on reliability was determined if an item was removed from the assessment.

Constructed-Response Items

Analyses for the constructed-response items included calculating the difficulty of each item and examining reliability of these items. Reliability analyses were conducted for constructed-response items in a fashion similar to the multiple-choice items. Item-total correlations were calculated for each item and reliability was estimated should the item be removed from the assessment. Likewise we estimated reliability for the complete set of items if all constructed-response questions were included on a single assessment.

Results

Below we present the results of our item analyses. However, we first discuss the issue of missing data and how we addressed this issue.

Missing Data

A major challenge in all research, particularly when developing measures, is missing data. It may be recalled that when designing the pilot test and the different forms that were administered to participants, we made some significant choices. First, we decided to include every item from the item pool in the pilot under the assumption that having some data on every item was more important than having no data on a significant number of items. Second, we counterbalanced sections of the test to guard against order and fatigue effects. Although these strategies were well thought out and had certain advantages, they did contribute to the amount of missing data.

In examining the raw responses, we identified two types of missing data. One occurred when an item was near the end of a section, and the participant simply did not have enough time to answer the item (Type A: *Did Not See*). The second type occurred when a participant could have answered the item but intentionally skipped the item for whatever reason (Type B: *Skipped*). Note that missing data were not systematically tracked; Types A and B were determined by exploring missing data trends. We presumed that the steady decrease of responses toward the end of each survey section signified that teachers were running out of time and were unable to complete their items.

The multiple-choice items suffered more than any other item type from the two forms of missing data described above. This outcome was unanticipated because of the safeguards we employed when designing the items and structuring the pilot test. For example, to determine the amount of time the multiple-choice items would take, we enlisted several research assistants to complete the items as they would for an actual examination. We timed them during this process and discovered that it required 0.5 minutes per item. To be conservative and account for the range of computer skills among respondents, we estimated for the actual survey that each multiple-choice item to take approximately 1 minute. Based on these calculations, and the test time allocated to the multiple-choice items (80 minutes for 88 and 83 items), we expected that most teachers would complete the items in these sections. However, data suggested that a large number of teachers ran out of time. Figure 1 below illustrates the downward trend in responses as teachers approached the end of the multiple-choice items. Notice how each section begins with over 200 teachers completing the items. However, over half of the teachers did not reach the end of the section with approximately 60 respondents completing the last item.

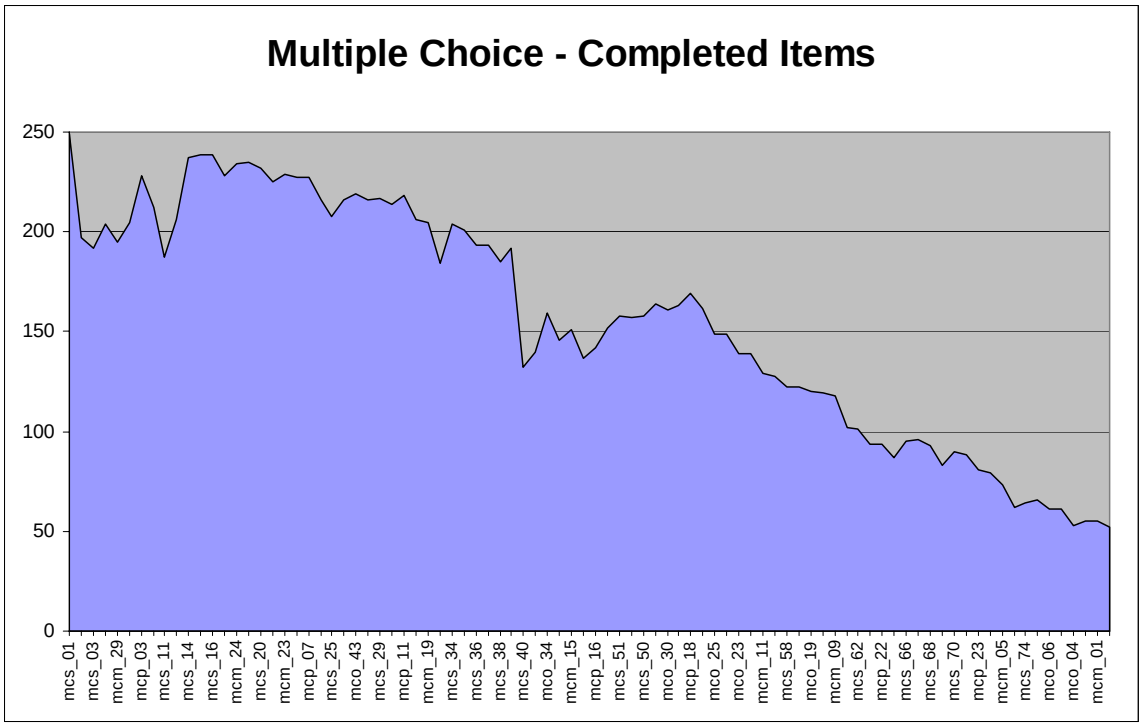

Figure 1. Multiple-choice item response rates by item order.

One possible explanation for the large number of skipped items was the lack of a "valued" incentive. Teachers who volunteered to participate in this pilot study were paid a reasonable amount, but not given any explicit enticements to respond to all items and there were no penalties for running out of time before completing each section. Alternatively, teachers may have been skipping items that they did not know how to answer. Given the nature of these items and the fact that they cover a wide range of testable content, many of the participants may have recognized that some items were beyond the scope of their knowledge. Thus, if a participant knew that s/he was not familiar with a particular domain, they may have seen no reason to expend sufficient effort on the item. This sizeable amount of unanticipated missing data on the multiple-choice items posed a significant challenge when examining reliability of a set of items which relies on complete data on each item in the analysis. Our strategy for dealing with this situation is discussed when report the reliability results.

Item Analysis Results

The following sections present the results of our item analyses for the 108 multiple-choice and 24 constructed-response items that we propose for the pre-service teacher knowledge assessment. For each item type, descriptive data are presented first followed by our reliability analysis. Once these results have been reviewed, we examine the extent to which the difficulty of the items varied as a function of participant experience. Because the assessment will be administered to pre-service teachers, we wanted to determine if performance on the items was a function of experience in the classroom.

Multiple Choice Items

Item Difficulty. Overall the difficulty analysis showed that multiple-choice items were moderate to high in difficulty (percent of respondents answering an item correctly). The average item difficulty across all 108 items was p=.53, with difficulty ranging from a low of p=.01, for one of the items designed to measure vocabulary, to p=.97, for one of the comprehension items. Table 7 presents additional information on item difficulty for the items broken down by each of the five NRP components. Referring to Table 7, item difficulty was similar across components.

Table 7. Summary of Multiple-Choice Item Characteristics

NRP Component	Number of items	N Range	Average Difficulty	Difficulty Range
Comprehension	27	52 -258	0.55	.16 -.97
Fluency	10	95 -229	0.46	.06 -.95
Phonemic Awareness	16	139 -273	0.51	.09 -.85
Phonics	22	90 - 254	0.57	.10 -.91
Vocabulary	33	53 – 238	0.53	.01 - .93
Total:	108			

Appendix A presents complete descriptive data for each of the multiple-choice items including the number of respondents, the item's difficulty, the answer key, and the distribution of responses across each item's response options. Referring to the appendix, there were 11 items that appeared to be miss keyed or had problems with the distractors. For example, for one of the fluency items (mcs_66) ***d*** was the correct answer, but 84% of the respondents selected ***b*** as the correct alternative. In such cases the keys were checked and verified to ensure the data were coded correctly. For the pre-service teacher assessment, it might be best not to include such items since they seem to be either too difficult or too confusing for the respondents.

Reliability. As described earlier, our reliability analysis focused on the internal consistency of items. Our expectation was that items developed to measure knowledge of the NRP should relate to one another. In other words, the 56 items on Version 1 should be internally consistent with one another as well as the 52 items on Version 2.

To determine reliability, we calculated alpha and item-total correlations for each version of the assessment. However, it will be recalled that missing data were most prevalent for the multiple-choice items. Furthermore missing data are most problematic when determining reliability because this analysis requires complete cases on the items of interest. Therefore, we treated missing data as incorrect for this round of our analyses. Although not completely desirable, this approach has been employed in other AIR high-stakes testing projects. Furthermore, treating missing data as wrong should only slightly enhance the item-total correlations and the alphas as opposed to significantly over estimating these values (AIR staff has conducted Monte Carlo studies testing this assumption). Nonetheless, because we had to employ this approach, we view these results as preliminary estimates.

Table 8 presents information on the range of item-total correlations and alpha for each version of the assessment. Appendix B presents item-total correlations for all of the multiple-choice items. Referring to Table 8, the alphas for each version of the

Table 8. Reliability and Item-total Correlations for the Multiple-Choice Items on each Version of the Assessment

	Number of items	N	Alpha	Item-Total Range
Version 1	56	283	0.73	-.06 - .51
Version 2	52	306	0.75	-.04 - .43

assessment exceed .7, which is reasonably high in magnitude. Because reliability is directly related to test length and the reliabilities in Table 8 are essentially based on half the number of items we would administer to assess pre-service teacher knowledge, we estimated the reliability of the proposed assessment by applying the Spearman-Brown formula to the reliability estimate for Version 1 (the lower value). This analysis indicated that the reliability for the entire set of multiple-choice items would be .84.

Constructed-Response Item Analysis

Item Difficulty. The proportion of individuals who answered an item correctly over the total number of individuals is used to calculate item difficulty of dichotomously scored items. However, since the constructed-response items were scored using 2, 1, and 0, for full credit, partial credit, and no credit, the conventional index of item difficulty was not used. Therefore, we created an index of item difficulty based on a procedure developed by the University of Iowa. This procedure yields an index of item difficulty that ranges from 0 (extremely hard) to 1 (extremely easy), which allows for comparison of difficulty levels to other item types.

Table 9. Summary of Constructed-Response Item Characteristics

NRP Component	Number of items	N Range	Average Difficulty	Difficulty Range
Comprehension	4	95 -177	0.54	.38 - .76
Fluency	5	90 -126	0.60	.43 - .92
Phonemic Awareness	3	49 -189	0.42	.25 - .61
Phonics	4	122 - 194	0.47	.32 - .70
Vocabulary	8	64 – 190	0.66	.51 - .94
Total:	24			

Overall the difficulty analysis showed that the constructed-response items were moderate to high in difficulty. The average item difficulty across all 24 items was p=.57, with difficulty ranging from a low of p=.25, for one of the items designed to measure phonemic awareness, to p=.94, for one of the vocabulary items. Table 9 presents additional information on item difficulty for the items broken down by each of the five NRP components. Appendix C presents complete descriptive data for each of the constructed-response items including the number of respondents, the item's difficulty, the average item score, and the percent agreement among raters.

Reliability. Similar to multiple-choice items, missing data on the constructed-response questions were treated as incorrect responses and item-total correlations and alphas for each version of the assessment were calculated. Table 10 reports the range of item-total correlations for each version of the survey while Appendix D reports complete item-level data and results.

Table 10. Reliability and Item-total Correlations for the Constructed-Response Items on each Version of the Assessment

	Number of items	N	Alpha	Item-Total Range
Version 1	12	283	0.54	-.06 - .49
Version 2	12	306	0.53	-.06 - .44

Referring to Table 10, the reliabilities for the constructed-response items were somewhat lower than desired (above .7). Similar to the multiple-choice items, we estimated the reliability for all 24 items using Spearman-Brown formula. This analysis indicated that the reliability for the entire set of constructed-response items would be .69.

One way to improve reliability of the constructed-response items is to remove the items with low item-total correlations. To test the effects of this strategy, we removed four items from Version 1 and three items from Version 2 whose item-total correlations were approximately zero. This analysis improved the alphas for the constructed-response items to .73 and .69 for Versions 1 and 2, respectively. Applying the Spearman Brown formula to the lower alpha (Version 2) produced a reliability estimate of .82 for the remaining 17 constructed-response items.

Effect of Teaching Experience

One important question that we wanted to answer is whether or not teaching experience affects performance on the items we propose for the teacher knowledge test. However, because the teachers in the pilot test were currently practicing teachers and the teacher assessment will be administered to pre-service teachers we could only indirectly answer this question. We were interested in this question because we had tried to develop items that assessed both declarative and procedural knowledge, and we reasoned that teachers who could draw on extensive experience in classroom settings might score higher on the test.

To explore the effects of experience, we created four groups of teachers based on the demographic data collected. The first group consisted of teachers who had three years or less teaching experience (N= 99), the second group consisted of teachers who had four to six years experience (N=114), the third group consisted of teachers who had seven to nine years experience (N=103), and the fourth group consisted of teachers who had 10 or more years experience (N=265). Next, we calculated the difficulty of the items for each of these experience subgroups to see if the item difficulty varied by subgroup. Table 11 reports the mean difficulty of the multiple-choice and constructed-response questions by subgroup. For the multiple-choice we also calculated these difficulty estimates by NRP component. Appendix E presents difficulty estimates for each

individual item by subgroup. Referring to the table and the appendix, item difficulty varied little as a function of teaching experience.

Table 11. Mean Item Difficulty by Experience

	Less than 3 years	4 to 6 years	7 to 9 years	10 or more years
Multiple Choice				
Comprehension	.55	.58	.56	.53
Fluency	.46	.47	.46	.45
Phonemic Awareness	.51	.49	.53	.52
Phonics	.53	.55	.59	.58
Vocabulary	.53	.52	.52	.53
Constructed-Response	.55	.57	.58	.56

To further explore whether or not experience affected performance on the items, we correlated the difficulty estimates across items for those teachers with three or fewer years of experience with each of the more experienced groups. While the averages in Table 11 indicate if the items are similar, the correlations presented in Table 12 indicate whether or not the items rank consistently with respect to difficulty across the different experience groups. In other words, the results in Table 12 show that the items which were easiest and hardest were the same regardless of experience.

Table 12. Correlations of Item Difficulty Estimates for Teachers with Less than Three Years Experience with the More Experienced Groups

	4 to 6 years	7 to 9 years	10 or more years
Multiple Choice			
Comprehension	.87	.92	.90
Fluency	.87	.95	.94
Phonemic Awareness	.88	.90	.92
Phonics	.94	.90	.90
Vocabulary	.84	.93	.92
Constructed-Response	.87	.88	.91

Combined, although only a proxy, the results presented here indicate that years of teaching experience is not related to performance on these items. However, we recognize that no pre-service teachers actually completed the items and therefore, how the items perform with a highly inexperienced sample may be different. We will rely on the TWG to provide us guidance as to whether or not, based on the data presented and the items themselves, the items are likely to perform differently when used to assess pre-service teacher knowledge of the NRP.

Summary

In summary, this report presents data on the performance of a set of multiple-choice and constructed-response items that were designed to assess teacher knowledge of the NRP. Data were collected on these items from 589 teachers as part of a larger item set that was pilot tested under the Instructional Processes Research and Development project sponsored by the National Center of Education Statistics. Although not the ideal approach for determining the performance of the items presented here, the pilot test results do provide a first look at how effective these items would be at assessing pre-service teacher knowledge of the NRP. Moreover, the results should aid the TWG in making a determination of the viability of using these items for the pre-service assessment.

To continue to develop a better understanding of the characteristics of these items, we are exploring the possibility of collecting additional data. Currently, the items are being pilot tested as part of AIR's Professional Development Impact project. Approximately 80 completed cases on these items should be available in January 2006 (though as with this effort these items are embedded in a larger pilot). We also would consider administering these items in a single assessment to a group of pre-service teachers, should the TWG be able to assist us in identifying such a sample. Although we are confident in our proposed assessment as is, we believe that it is important to explore any additional options for verifying the reliability and validity of this assessment.

Appendix A. Multiple-Choice Descriptive Data

							Response Percentages			
Item No.	**NRP Component**	**Version**	**N**	**Response Rate**	**Difficulty**	**Correct Answer**	**1**	**2**	**3**	**4**
mcs_17	Comprehension	2	258	95%	0.47	1	**47**	*5*	*2*	45
mcs_31	Comprehension	1	206	83%	0.48	4	15	19	18	**48**
mcs_44	Comprehension	1	137	55%	0.97	4	*1*	*2*	*1*	**97**
mcs_45	Comprehension	2	173	63%	0.53	1	**53**	*3*	11	33
mcs_46	Comprehension	1	152	61%	0.86	3	*5*	7	**86**	*2*
mcs_47	Comprehension	2	197	72%	0.45	3	*3*	45	**45**	7
mcs_48	Comprehension	2	190	70%	0.66	2	11	**66**	10	13
mco_01	Comprehension	2	63	23%	0.43	3	13	6	**43**	38
mco_02	Comprehension	1	52	21%	0.60	4	*2*	31	8	**60**
mco_03	Comprehension	2	64	23%	0.41	2	19	**41**	30	11
mco_14	Comprehension	2	109	40%	0.82	4	*0*	7	11	**82**
mco_15	Comprehension	1	87	35%	0.26	4	21	13	40	**26**
mco_16	Comprehension	2	129	47%	0.64	2	14	**64**	*3*	19
mco_17	Comprehension	1	119	48%	0.54	3	*1*	28	**54**	18
mco_18	Comprehension	2	135	49%	0.77	4	13	*5*	*4*	**77**
mco_19	Comprehension	1	120	48%	0.60	4	*5*	*4*	31	**60**
mco_27	Comprehension	1	157	63%	0.54	3	24	8	**54**	15
mco_29	Comprehension	1	164	66%	0.87	2	9	**87**	*3*	*1*
mco_30	Comprehension	1	161	65%	0.54	4	*1*	*3*	42	**54**
mco_31	Comprehension	2	174	64%	0.21	1	**21**	47	*1*	31
mco_32	Comprehension	2	187	68%	0.62	1	**62**	*3*	24	12
mco_33	Comprehension	1	140	56%	0.50	3	8	38	**50**	*4*
mco_34	Comprehension	1	159	64%	0.83	1	**83**	15	*0*	*2*
mco_35	Comprehension	2	157	58%	0.20	4	21	50	9	**20**
mco_45	Comprehension	1	216	87%	0.16	2	25	**16**	20	39
mco_46	Comprehension	2	238	87%	0.38	3	30	27	**38**	*5*
mco_47	Comprehension	1	225	90%	0.60	4	28	*3*	8	**60**

							Response Percentages			
Item No.	**NRP Component**	**Version**	**N**	**Response Rate**	**Difficulty**	**Correct Answer**	**1**	**2**	**3**	**4**
mcs_32	Fluency	2	229	84%	0.66	2	29	**66**	6	*0*
mcs_34	Fluency	1	204	82%	0.37	2	61	**37**	*1*	*2*
mcs_35	Fluency	2	220	81%	0.31	3	33	31	**31**	6
mcs_36	Fluency	1	193	78%	0.40	4	*1*	21	39	**40**
mcs_50	Fluency	1	158	63%	0.54	2	40	**54**	*2*	*4*
mcs_64	Fluency	2	141	52%	0.14	2	25	**14**	10	52
mcs_65	Fluency	2	110	40%	0.55	4	10	9	26	**55**
mcs_66	Fluency	1	95	38%	0.06	4	8	84	*2*	**6**
mco_20	Fluency	2	160	59%	0.95	3	*1*	*3*	**95**	*1*
mco_36	Fluency	2	215	79%	0.62	1	**62**	8	9	21

							Response Percentages			
Item No.	**NRP Component**	**Version**	**N**	**Response Rate**	**Difficulty**	**Correct Answer**	**1**	**2**	**3**	**4**
mcs_01	Phonemic Awareness	1	249	100%	0.43	1	**43**	35	10	12
mcs_02	Phonemic Awareness	2	273	100%	0.74	3	9	16	**74**	*1*
mcs_03	Phonemic Awareness	1	192	77%	0.20	4	28	27	26	**20**
mcs_04	Phonemic Awareness	2	270	99%	0.48	3	49	*1*	**48**	*3*
mcs_05	Phonemic Awareness	1	204	82%	0.35	1	**35**	6	33	26
mcs_13	Phonemic Awareness	2	166	61%	0.85	3	7	7	**85**	*1*
mcs_14	Phonemic Awareness	1	237	95%	0.33	4	52	14	*0*	**33**
mcs_15	Phonemic Awareness	2	249	91%	0.79	2	6	**79**	14	*1*
mcs_16	Phonemic Awareness	1	239	96%	0.09	2	6	**9**	84	*1*
mcs_55	Phonemic Awareness	2	183	67%	0.81	3	*2*	*2*	**81**	16
mcs_56	Phonemic Awareness	1	139	56%	0.54	3	30	6	**54**	11
mco_23	Phonemic Awareness	1	139	56%	0.25	4	*1*	68	6	**25**
mco_24	Phonemic Awareness	2	161	59%	0.67	3	12	11	**67**	11
mco_25	Phonemic Awareness	1	149	60%	0.63	3	7	16	**63**	13
mco_26	Phonemic Awareness	2	193	71%	0.73	4	8	*1*	19	**73**
mco_43	Phonemic Awareness	1	219	88%	0.34	4	32	16	18	**34**

							Response Percentages			
Item No.	**NRP Component**	**Version**	**N**	**Response Rate**	**Difficulty**	**Correct Answer**	**1**	**2**	**3**	**4**
mcs_06	Phonics	2	180	66%	0.75	4	24	*0*	*1*	**75**
mcs_18	Phonics	1	235	94%	0.84	2	*1*	**84**	11	*4*
mcs_19	Phonics	2	253	93%	0.78	4	*0*	*0*	22	**78**
mcs_20	Phonics	1	232	93%	0.89	2	*3*	**89**	6	*2*
mcs_21	Phonics	2	230	84%	0.39	1	**39**	31	16	14
mcs_22	Phonics	1	227	91%	0.53	2	12	**53**	32	*4*
mcs_24	Phonics	2	252	92%	0.90	3	*2*	*4*	**90**	*5*
mcs_25	Phonics	1	208	84%	0.59	2	16	**59**	18	7
mcs_26	Phonics	2	239	88%	0.74	1	**74**	*2*	8	16
mcs_37	Phonics	2	216	79%	0.71	3	*4*	12	**71**	13
mcs_38	Phonics	1	185	74%	0.21	2	64	**21**	10	*5*
mcs_39	Phonics	2	213	78%	0.10	4	16	32	43	**10**
mcs_51	Phonics	1	158	63%	0.76	1	**76**	19	*4*	*1*
mcs_52	Phonics	1	162	65%	0.53	3	19	*4*	**53**	24
mcs_68	Phonics	1	93	37%	0.91	4	*3*	*4*	*1*	**91**
mcs_69	Phonics	2	98	36%	0.37	3	32	16	**37**	15
mcs_70	Phonics	1	90	36%	0.36	4	30	8	27	**36**
mco_37	Phonics	1	193	78%	0.78	4	6	16	*1*	**78**
mco_38	Phonics	2	210	77%	0.47	1	**47**	9	28	17
mco_39	Phonics	1	184	74%	0.25	1	**25**	33	*5*	36
mco_40	Phonics	2	205	75%	0.33	3	7	25	**33**	36
mco_48	Phonics	2	254	93%	0.31	3	39	7	**31**	23

							Response Percentages			
Item No.	**NRP Component**	**Version**	**N**	**Response Rate**	**Difficulty**	**Correct Answer**	**1**	**2**	**3**	**4**
mcs_07	Vocabulary	1	205	82%	0.49	1	**49**	6	11	34
mcs_08	Vocabulary	2	165	60%	0.01	1	***1***	7	90	*2*
mcs_09	Vocabulary	1	212	85%	0.20	4	60	8	12	**20**
mcs_10	Vocabulary	2	168	62%	0.78	3	9	13	**78**	*0*
mcs_11	Vocabulary	1	187	75%	0.82	2	*2*	**82**	*2*	14
mcs_27	Vocabulary	1	216	87%	0.60	3	7	*4*	**60**	29
mcs_29	Vocabulary	1	217	87%	0.83	2	6	**83**	*2*	10
mcs_30	Vocabulary	2	238	87%	0.93	4	*1*	*0*	6	**93**
mcs_40	Vocabulary	1	132	53%	0.72	2	*1*	**72**	21	6
mcs_41	Vocabulary	2	169	62%	0.53	1	**53**	6	29	12
mcs_42	Vocabulary	1	146	59%	0.32	3	32	26	**32**	10
mcs_43	Vocabulary	2	150	55%	0.89	3	7	*1*	**89**	*3*
mcs_53	Vocabulary	2	184	67%	0.75	2	*2*	**75**	*2*	21
mcs_54	Vocabulary	1	149	60%	0.34	3	9	11	**34**	54
mcs_71	Vocabulary	2	97	36%	0.47	3	13	34	**47**	*5*
mcs_72	Vocabulary	1	79	32%	0.24	4	10	24	42	**24**
mcs_73	Vocabulary	2	94	34%	0.13	2	15	**13**	70	*2*
mcs_74	Vocabulary	1	64	26%	0.48	4	17	31	*3*	**48**
mcs_75	Vocabulary	2	72	26%	0.69	3	*1*	26	**69**	*3*
mcs_76	Vocabulary	1	66	27%	0.77	2	*2*	**77**	21	*0*
mco_04	Vocabulary	1	53	21%	0.15	4	13	30	42	**15**
mco_05	Vocabulary	2	66	24%	0.67	3	20	*3*	**67**	11
mco_06	Vocabulary	1	61	24%	0.46	2	36	**46**	13	*5*
mco_07	Vocabulary	2	73	27%	0.52	4	*4*	11	33	**52**
mco_08	Vocabulary	1	62	25%	0.23	3	*5*	16	**23**	57
mco_09	Vocabulary	2	96	35%	0.55	3	25	*2*	**55**	18
mco_10	Vocabulary	1	88	35%	0.53	1	**53**	10	32	*5*

							Response Percentages			
Item No.	**NRP Component**	**Version**	**N**	**Response Rate**	**Difficulty**	**Correct Answer**	**1**	**2**	**3**	**4**
mco_11	Vocabulary	2	94	34%	0.27	3	17	27	**27**	30
mco_12	Vocabulary	1	83	33%	0.40	4	28	18	15	**40**
mco_21	Vocabulary	1	128	51%	0.48	1	**48**	9	41	*2*
mco_22	Vocabulary	2	180	66%	0.82	1	**82**	13	*2*	*3*
mco_41	Vocabulary	1	214	86%	0.67	3	16	*4*	**67**	13
mco_42	Vocabulary	2	237	87%	0.73	3	13	*3*	**73**	11

Appendix B. Multiple-Choice Item-Total Correlations

Multiple Choice Item-Total Statistics for Version 1

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Item-Total Correlation	Alpha if Item Deleted
mco_02_cscore	15.7915	34.698	0.057	0.734
mco_04_cscore	15.8728	35.083	-0.058	0.735
mco_06_cscore	15.7986	34.771	0.039	0.735
mco_08_cscore	15.8481	35.101	-0.058	0.736
mco_10_cscore	15.7350	34.593	0.062	0.735
mco_12_cscore	15.7845	34.914	-0.004	0.736
mco_15_cscore	15.8163	34.739	0.058	0.734
mco_17_cscore	15.6714	34.321	0.103	0.734
mco_19_cscore	15.6466	33.875	0.185	0.730
mco_21_cscore	15.6855	34.060	0.162	0.731
mco_23_cscore	15.7774	34.464	0.112	0.733
mco_25_cscore	15.5689	33.544	0.227	0.728
mco_27_cscore	15.6007	32.702	0.399	0.720
mco_29_cscore	15.3922	32.934	0.316	0.724
mco_30_cscore	15.5936	32.852	0.367	0.722
mco_33_cscore	15.6502	33.427	0.277	0.726
mco_34_cscore	15.4311	32.126	0.463	0.716
mco_37_cscore	15.3640	32.849	0.332	0.723
mco_39_cscore	15.7350	33.777	0.253	0.728
mco_41_cscore	15.3958	33.587	0.200	0.730
mco_43_cscore	15.6360	33.949	0.167	0.731
mco_45_cscore	15.7739	34.743	0.038	0.735
mco_47_cscore	15.4205	33.819	0.160	0.732
mcs_01_cscore	15.5159	33.790	0.172	0.731
mcs_03_cscore	15.7597	34.722	0.039	0.735
mcs_05_cscore	15.6466	34.017	0.157	0.731
mcs_07_cscore	15.5406	34.469	0.053	0.736
mcs_09_cscore	15.7456	34.871	0.000	0.737
mcs_11_cscore	15.3569	34.081	0.115	0.734
mcs_14_cscore	15.6219	33.775	0.197	0.730
mcs_16_cscore	15.8233	34.451	0.156	0.731
mcs_18_cscore	15.2014	33.793	0.186	0.730
mcs_20_cscore	15.1661	33.990	0.157	0.731
mcs_22_cscore	15.4735	34.179	0.099	0.734
mcs_25_cscore	15.4664	34.328	0.073	0.736
mcs_27_cscore	15.4452	34.312	0.075	0.736
mcs_29_cscore	15.2650	33.408	0.243	0.728
mcs_31_cscore	15.5512	33.596	0.213	0.729
mcs_34_cscore	15.6325	34.056	0.145	0.732
mcs_36_cscore	15.6219	34.016	0.150	0.732
mcs_38_cscore	15.7633	34.344	0.135	0.732
mcs_40_cscore	15.5618	33.226	0.284	0.726
mcs_42_cscore	15.7314	34.162	0.160	0.731
mcs_44_cscore	15.4276	32.182	0.453	0.717
mcs_46_cscore	15.4346	31.885	0.508	0.714

Multiple Choice Item-Total Statistics for Version 1

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Item-Total Correlation	Alpha if Item Deleted
mcs_50_cscore	15.5972	32.568	0.424	0.719
mcs_51_cscore	15.4735	32.371	0.423	0.719
mcs_52_cscore	15.5972	33.163	0.307	0.725
mcs_54_cscore	15.7244	33.548	0.298	0.726
mcs_56_cscore	15.6325	33.340	0.287	0.726
mcs_66_cscore	15.8763	34.960	0.011	0.734
mcs_68_cscore	15.5972	34.348	0.082	0.735
mcs_70_cscore	15.7845	34.695	0.055	0.734
mcs_72_cscore	15.8304	34.872	0.022	0.735
mcs_74_cscore	15.7915	34.627	0.077	0.734
mcs_76_cscore	15.7173	34.714	0.030	0.736

Multiple Choice Item-Total Statistics for Version 2

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Item-Total Correlation	Alpha if Item Deleted
mco_01_cscore	17.1013	35.560	0.180	0.745
mco_03_cscore	17.1046	35.484	0.207	0.744
mco_05_cscore	17.0458	34.982	0.276	0.741
mco_07_cscore	17.0654	35.471	0.171	0.745
mco_09_cscore	17.0163	34.777	0.298	0.740
mco_11_cscore	17.1078	35.218	0.295	0.742
mco_14_cscore	16.8987	34.445	0.300	0.740
mco_16_cscore	16.9216	34.669	0.265	0.741
mco_18_cscore	16.8497	34.115	0.345	0.737
mco_20_cscore	16.6928	33.702	0.395	0.735
mco_22_cscore	16.7059	33.487	0.434	0.733
mco_24_cscore	16.8399	33.761	0.408	0.734
mco_26_cscore	16.7320	33.856	0.370	0.736
mco_31_cscore	17.0719	35.910	0.062	0.748
mco_32_cscore	16.8137	34.736	0.224	0.743
mco_35_cscore	17.0850	35.691	0.128	0.746
mco_36_cscore	16.7549	34.573	0.245	0.742
mco_38_cscore	16.8693	34.763	0.231	0.742
mco_40_cscore	16.9706	36.061	0.004	0.751
mco_42_cscore	16.6209	34.512	0.256	0.741
mco_46_cscore	16.8922	35.690	0.065	0.749
mco_48_cscore	16.9314	35.881	0.034	0.750
mcs_02_cscore	16.5294	35.699	0.058	0.750
mcs_04_cscore	16.7680	35.202	0.137	0.747
mcs_06_cscore	16.7484	35.710	0.050	0.751
mcs_08_cscore	17.1830	36.176	0.072	0.747
mcs_10_cscore	16.7614	35.035	0.166	0.745
mcs_13_cscore	16.7288	35.352	0.110	0.748
mcs_15_cscore	16.5458	34.983	0.183	0.745
mcs_17_cscore	16.7908	35.386	0.107	0.748
mcs_19_cscore	16.5490	35.199	0.144	0.746
mcs_21_cscore	16.8987	35.954	0.017	0.751
mcs_24_cscore	16.4510	35.114	0.181	0.744
mcs_26_cscore	16.6111	34.907	0.188	0.744
mcs_30_cscore	16.4641	35.030	0.193	0.744
mcs_32_cscore	16.6993	35.372	0.106	0.748
mcs_35_cscore	16.9706	35.202	0.179	0.745
mcs_37_cscore	16.6895	34.949	0.178	0.745
mcs_39_cscore	17.1176	36.301	-0.037	0.750
mcs_41_cscore	16.8954	35.425	0.114	0.747
mcs_43_cscore	16.7516	34.974	0.175	0.745
mcs_45_cscore	16.8889	34.447	0.296	0.740
mcs_47_cscore	16.9020	34.948	0.205	0.744
mcs_48_cscore	16.7778	33.996	0.350	0.737

Multiple Choice Item-Total Statistics for Version 2

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Item-Total Correlation	Alpha if Item Deleted
mcs_53_cscore	16.7386	33.676	0.403	0.734
mcs_55_cscore	16.7059	33.579	0.418	0.734
mcs_64_cscore	17.1275	36.000	0.067	0.747
mcs_65_cscore	16.9935	35.056	0.221	0.743
mcs_69_cscore	17.0719	35.234	0.239	0.743
mcs_71_cscore	17.0392	35.330	0.187	0.744
mcs_73_cscore	17.1503	35.840	0.161	0.746
mcs_75_cscore	17.0261	34.885	0.281	0.741

Appendix C. Constructed-Response Descriptive Data

Item No.	NRP Component	Version	N	Difficulty	Average score	Percent agreement among raters
sao_11	Comprehension	1	118	0.38	0.76	80.93
sao_13	Comprehension	1	107	0.45	0.90	82.24
sao_06	Comprehension	2	95	0.76	1.52	91.58
sas_05_01	Comprehension	2	177	0.57	1.14	86.35
sao_08	Fluency	1	93	0.52	1.03	85.48
sas_14	Fluency	1	114	0.55	1.09	82.46
sas_22	Fluency	1	91	0.92	1.84	93.41
sas_13_01	Fluency	2	90	0.58	1.17	86.11
sas_15	Fluency	2	126	0.43	0.87	75.66
sao_18_02	Phonemic Awareness	2	120	0.61	1.21	90.00
sas_04	Phonemic Awareness	1	189	0.25	0.51	78.04
sas_24	Phonemic Awareness	2	49	0.40	0.80	78.57
sao_17	Phonics	1	194	0.32	0.65	89.86
sas_09	Phonics	1	190	0.52	1.04	90.26
sas_16	Phonics	1	135	0.70	1.41	87.04
sas_08	Phonics	2	122	0.32	0.64	93.44
sao_09	Vocabulary	2	134	0.64	1.28	84.33
sas_18	Vocabulary	1	111	0.67	1.34	90.54
sas_20	Vocabulary	1	86	0.52	1.05	88.37
sas_26	Vocabulary	1	64	0.55	1.09	79.69
sas_01_01	Vocabulary	2	190	0.85	1.69	90.00
sas_11	Vocabulary	2	166	0.60	1.20	84.04
sas_19	Vocabulary	2	122	0.51	1.03	79.51
sas_21	Vocabulary	2	117	0.94	1.88	93.59

Appendix D. Constructed-Response Item-Total Correlations

Item No.	NRP Component	Version	N	Item-Total Correlation	Alpha if Item Deleted
sao_11	Comprehension	1	283	.37	.48
sao_13	Comprehension	1	283	.21	.51
sao_08	Fluency	1	283	.37	.48
sas_14*	Fluency	1	283	.02	.56
sas_22	Fluency	1	283	.41	.45
sas_04*	Phonemic Awareness	1	283	-.07	.56
sao_17*	Phonics	1	283	.06	.55
sas_09*	Phonics	1	283	-.06	.58
sas_16	Phonics	1	283	.11	.54
sas_18	Vocabulary	1	283	.41	.46

Note: (*) indicates items that were deleted for subsequent reliability analysis.

Item No.	NRP Component	Version	N	Item-Total Correlation	Alpha if Item Deleted
sao_06	Comprehension	2	306	.44	.44
sas_05_01	Comprehension	2	306	.28	.49
sas_13_01	Fluency	2	306	.28	.49
sas_15	Fluency	2	306	.12	.53
sao_18_02*	Phonemic Awareness	2	306	-.06	.58
sas_24	Phonemic Awareness	2	306	.36	.49
sas_08*	Phonics	2	306	-.06	.56
sao_09	Vocabulary	2	306	.36	.47
sas_01_01*	Vocabulary	2	306	-.05	.59
sas_11	Vocabulary	2	306	.10	.54
sas_19	Vocabulary	2	306	.47	.45
sas_21	Vocabulary	2	306	.44	.43

Note: (*) indicates items that were deleted for subsequent reliability analysis.

Appendix E. Item Difficulties by Teacher Experience

Multiple – Choice Item No.	NRP Component	Version	Full Sample	3 years or fewer	4 to 6 years	7 to 9 years	10 or more years
mco_02	Comprehension	1	0.60	0.44	0.70	0.50	0.65
mco_15	Comprehension	1	0.26	0.29	0.13	0.39	0.25
mco_17	Comprehension	1	0.54	0.57	0.65	0.58	0.45
mco_19	Comprehension	1	0.60	0.55	0.68	0.60	0.58
mco_27	Comprehension	1	0.54	0.52	0.64	0.46	0.52
mco_29	Comprehension	1	0.87	0.93	0.81	0.88	0.88
mco_30	Comprehension	1	0.54	0.55	0.49	0.56	0.55
mco_33	Comprehension	1	0.50	0.58	0.59	0.57	0.40
mco_34	Comprehension	1	0.83	0.74	0.86	0.84	0.85
mco_45	Comprehension	1	0.16	0.15	0.25	0.14	0.14
mco_47	Comprehension	1	0.60	0.73	0.49	0.60	0.61
mcs_31	Comprehension	1	0.48	0.45	0.43	0.44	0.51
mcs_44	Comprehension	1	0.97	1.00	0.97	1.00	0.95
mcs_46	Comprehension	1	0.86	0.81	0.88	0.92	0.85
mco_01	Comprehension	2	0.43	0.43	0.56	0.33	0.37
mco_03	Comprehension	2	0.41	0.29	0.53	0.47	0.33
mco_14	Comprehension	2	0.82	0.85	0.89	0.75	0.79
mco_16	Comprehension	2	0.64	0.78	0.70	0.66	0.52
mco_18	Comprehension	2	0.77	0.85	0.71	0.80	0.74
mco_31	Comprehension	2	0.21	0.24	0.26	0.18	0.18
mco_32	Comprehension	2	0.62	0.59	0.61	0.63	0.62
mco_35	Comprehension	2	0.20	0.16	0.14	0.30	0.21
mco_46	Comprehension	2	0.38	0.34	0.30	0.42	0.42
mcs_17	Comprehension	2	0.47	0.33	0.40	0.54	0.54
mcs_45	Comprehension	2	0.53	0.56	0.56	0.58	0.48
mcs_47	Comprehension	2	0.45	0.38	0.56	0.49	0.40
mcs_48	Comprehension	2	0.66	0.68	0.76	0.59	0.64

Multiple – Choice Item No.	NRP Component	Version	Full Sample	3 years or fewer	4 to 6 years	7 to 9 years	10 or more years
mcs_34	Fluency	1	0.37	0.27	0.40	0.42	0.37
mcs_36	Fluency	1	0.40	0.44	0.39	0.27	0.44
mcs_50	Fluency	1	0.54	0.47	0.64	0.52	0.53
mcs_66	Fluency	1	0.06	0.06	0.17	0.05	0.03
mco_20	Fluency	2	0.95	1.00	0.97	0.97	0.90
mco_36	Fluency	2	0.62	0.62	0.53	0.59	0.68
mcs_32	Fluency	2	0.66	0.58	0.74	0.60	0.68
mcs_35	Fluency	2	0.31	0.37	0.33	0.28	0.28
mcs_64	Fluency	2	0.14	0.11	0.13	0.19	0.12
mcs_65	Fluency	2	0.55	0.68	0.39	0.71	0.49

Multiple – Choice Item No.	NRP Component	Version	Full Sample	3 years or fewer	4 to 6 years	7 to 9 years	10 or more years
mco_23	Phonemic Awareness	1	0.25	0.39	0.17	0.40	0.16
mco_25	Phonemic Awareness	1	0.63	0.60	0.48	0.76	0.66
mco_43	Phonemic Awareness	1	0.34	0.32	0.40	0.30	0.33
mcs_01	Phonemic Awareness	1	0.43	0.44	0.39	0.35	0.47
mcs_03	Phonemic Awareness	1	0.20	0.13	0.16	0.25	0.23
mcs_05	Phonemic Awareness	1	0.35	0.38	0.26	0.27	0.41
mcs_14	Phonemic Awareness	1	0.33	0.26	0.35	0.31	0.36
mcs_16	Phonemic Awareness	1	0.09	0.03	0.08	0.13	0.10
mcs_56	Phonemic Awareness	1	0.54	0.54	0.43	0.65	0.54
mco_24	Phonemic Awareness	2	0.67	0.61	0.84	0.69	0.60
mco_26	Phonemic Awareness	2	0.73	0.90	0.72	0.68	0.66
mcs_02	Phonemic Awareness	2	0.74	0.76	0.75	0.69	0.75
mcs_04	Phonemic Awareness	2	0.48	0.42	0.56	0.49	0.45
mcs_13	Phonemic Awareness	2	0.85	0.77	0.83	0.92	0.86
mcs_15	Phonemic Awareness	2	0.79	0.76	0.76	0.82	0.80
mcs_55	Phonemic Awareness	2	0.81	0.84	0.72	0.70	0.90

Multiple – Choice Item No.	NRP Component	Version	Full Sample	3 years or fewer	4 to 6 years	7 to 9 years	10 or more years
mco_37	Phonics	1	0.78	0.84	0.83	0.88	0.72
mco_39	Phonics	1	0.25	0.26	0.30	0.35	0.20
mcs_18	Phonics	1	0.84	0.74	0.81	0.81	0.89
mcs_20	Phonics	1	0.89	0.82	0.87	0.97	0.91
mcs_22	Phonics	1	0.53	0.31	0.59	0.57	0.57
mcs_25	Phonics	1	0.59	0.50	0.54	0.57	0.63
mcs_38	Phonics	1	0.21	0.27	0.15	0.17	0.22
mcs_51	Phonics	1	0.76	0.81	0.81	0.80	0.70
mcs_52	Phonics	1	0.53	0.55	0.59	0.45	0.52
mcs_68	Phonics	1	0.91	0.89	0.94	0.95	0.89
mcs_70	Phonics	1	0.36	0.33	0.24	0.35	0.43
mco_38	Phonics	2	0.47	0.51	0.49	0.43	0.46
mco_40	Phonics	2	0.33	0.39	0.29	0.27	0.35
mco_48	Phonics	2	0.31	0.22	0.37	0.33	0.31
mcs_06	Phonics	2	0.75	0.77	0.75	0.75	0.74
mcs_19	Phonics	2	0.78	0.64	0.67	0.91	0.81
mcs_21	Phonics	2	0.39	0.39	0.30	0.42	0.41
mcs_24	Phonics	2	0.90	0.86	0.88	0.85	0.94
mcs_26	Phonics	2	0.74	0.59	0.67	0.78	0.82
mcs_37	Phonics	2	0.71	0.65	0.69	0.74	0.73
mcs_39	Phonics	2	0.10	0.13	0.07	0.10	0.11
mcs_69	Phonics	2	0.37	0.21	0.30	0.48	0.43

Multiple – Choice Item No.	NRP Component	Version	Full Sample	3 years or fewer	4 to 6 years	7 to 9 years	10 or more years
mco_04	Vocabulary	1	0.15	0.09	0.11	0.11	0.21
mco_06	Vocabulary	1	0.46	0.50	0.15	0.64	0.52
mco_08	Vocabulary	1	0.23	0.23	0.18	0.23	0.24
mco_10	Vocabulary	1	0.53	0.50	0.61	0.50	0.53
mco_12	Vocabulary	1	0.40	0.41	0.36	0.44	0.38
mco_21	Vocabulary	1	0.48	0.36	0.46	0.38	0.57
mco_41	Vocabulary	1	0.67	0.61	0.84	0.59	0.64
mcs_07	Vocabulary	1	0.49	0.38	0.44	0.45	0.56
mcs_09	Vocabulary	1	0.20	0.11	0.18	0.14	0.26
mcs_11	Vocabulary	1	0.82	0.83	0.86	0.81	0.80
mcs_27	Vocabulary	1	0.60	0.67	0.61	0.55	0.58
mcs_29	Vocabulary	1	0.83	0.69	0.76	0.86	0.88
mcs_40	Vocabulary	1	0.72	0.81	0.67	0.72	0.71
mcs_42	Vocabulary	1	0.32	0.30	0.34	0.24	0.34
mcs_54	Vocabulary	1	0.34	0.30	0.44	0.30	0.32
mcs_72	Vocabulary	1	0.24	0.29	0.20	0.29	0.21
mcs_74	Vocabulary	1	0.48	0.54	0.21	0.58	0.56
mcs_76	Vocabulary	1	0.77	0.86	0.75	0.73	0.76
mco_05	Vocabulary	2	0.67	0.58	0.82	0.65	0.60
mco_07	Vocabulary	2	0.52	0.64	0.56	0.41	0.50
mco_09	Vocabulary	2	0.55	0.50	0.57	0.59	0.54
mco_11	Vocabulary	2	0.27	0.11	0.41	0.23	0.28
mco_22	Vocabulary	2	0.82	0.82	0.87	0.74	0.85
mco_42	Vocabulary	2	0.73	0.68	0.72	0.81	0.72
mcs_08	Vocabulary	2	0.01	0.00	0.06	0.00	0.00
mcs_10	Vocabulary	2	0.78	0.77	0.85	0.74	0.77
mcs_30	Vocabulary	2	0.93	0.88	0.93	0.94	0.95
mcs_41	Vocabulary	2	0.53	0.66	0.38	0.57	0.51

Multiple – Choice Item No.	NRP Component	Version	Full Sample	3 years or fewer	4 to 6 years	7 to 9 years	10 or more years
mcs_43	Vocabulary	2	0.89	0.97	0.84	0.97	0.84
mcs_53	Vocabulary	2	0.75	0.74	0.79	0.70	0.75
mcs_71	Vocabulary	2	0.47	0.67	0.43	0.38	0.46
mcs_73	Vocabulary	2	0.13	0.17	0.11	0.14	0.11
mcs_75	Vocabulary	2	0.69	0.71	0.67	0.78	0.64

Constructed-Response Item No.	NRP Component	Version	Full Sample Difficulty	3 years or fewer	4 to 6 years	7 to 9 years	10 or more years
sao_11	Comprehension	1	0.38	0.39	0.41	0.40	0.35
sao_13	Comprehension	1	0.45	0.52	0.50	0.50	0.38
sas_05_01	Comprehension	2	0.57	0.55	0.61	0.62	0.52
sao_06	Comprehension	2	0.76	0.81	0.79	0.70	0.73
sas_14	Fluency	1	0.55	0.54	0.38	0.61	0.59
sas_22	Fluency	1	0.92	0.83	0.92	0.98	0.92
sao_08	Fluency	1	0.52	0.56	0.47	0.55	0.49
sas_13_01	Fluency	2	0.58	0.47	0.56	0.56	0.66
sas_15	Fluency	2	0.43	0.33	0.48	0.44	0.45
sas_04	Phonemic Awareness	1	0.25	0.21	0.27	0.18	0.28
sas_24	Phonemic Awareness	2	0.40	0.32	0.43	0.48	0.37
sao_18_02	Phonemic Awareness	2	0.61	0.55	0.58	0.61	0.63
sas_09	Phonics	1	0.52	0.51	0.49	0.57	0.53
sas_16	Phonics	1	0.70	0.67	0.84	0.76	0.65
sao_17	Phonics	1	0.32	0.32	0.32	0.26	0.35
sas_08	Phonics	2	0.32	0.24	0.31	0.31	0.35
sas_18	Vocabulary	1	0.67	0.62	0.73	0.66	0.66
sas_20	Vocabulary	1	0.52	0.40	0.61	0.67	0.47
sas_26	Vocabulary	1	0.55	0.62	0.57	0.51	0.52
sas_01_01	Vocabulary	2	0.85	0.89	0.79	0.87	0.85
sas_11	Vocabulary	2	0.60	0.69	0.56	0.59	0.60
sas_19	Vocabulary	2	0.51	0.62	0.49	0.48	0.49
sas_21	Vocabulary	2	0.94	0.93	0.94	0.93	0.95
sao_09	Vocabulary	2	0.64	0.62	0.66	0.66	0.63