

Developing Achievement Levels on the 1998 National Assessment of Educational Progress in Civics: Validation Research

Susan Cooper Loomis
ACT, Inc.

December 2000

Developing Achievement Levels on the 1998 National Assessment of Educational Progress in Civics: Validation Research

Susan Cooper Loomis
ACT, Inc.

December 2000

The work for this report was conducted by ACT, Inc. under contract ZA97001001 with the National Assessment Governing Board.

Copyright © 2000 by ACT, Inc. All rights reserved.

Table of Contents

Civics Item Classification Study: 1998 Civics NAEP.....	1-1
Executive Summary	1-1
Background Information about the Study	1-3
Panelists	1-4
Materials	1-4
Advance Materials	1-4
Materials for the Study	1-4
Training for the Classification Task.....	1-5
The Classification Process	1-5
Data Analysis and Reporting	1-6
Agreement Among Teachers	1-6
Item Difficulty and Classification Level	1-7
Analyses Using Different Response Probability Criteria	1-8
Analysis of the Effect of Using a Correction for Guessing	1-8
Analysis of IRT Pseudo-Guessing Parameters for Civics Items.....	1-9
Summary.....	1-9
Reference	1-11
The 1998 Civics NAEP Validation Study of Achievement Levels	2-1
Executive Summary	2-1
Background of the Study	2-3
Overview of the Current Study	2-4
The Assessment Form Used for This Study	2-5
Selection of Item Blocks	2-5
Construction of the Forms.....	2-9
Planning and Logistics.....	2-9
Test Administration	2-10
Participation in the Study	2-11
Teacher Panelists and Their Schools	2-11
Students and Their Schools	2-12
Computation of Student Scores	2-15
The Study.....	2-16
Implementation of the Similarities Classification Study	2-17
Classification of the Overall Level of Civics Knowledge and Skills.....	2-17
An On-Site Change in the Study Design: Round 2 of Classifying the Overall Level of Civics Knowledge and Skills	2-17
Classification of Expected Student Performance on the Special Civics NAEP	2-18
Findings of the Similarities Classification Study	2-19
Overall Knowledge and Skill in Civics.....	2-19
Classification of Students' Expected Performance on the Special Form of NAEP	2-21
Summary of Findings	2-24
Booklet Classification Study.....	2-24
Selection of Booklets for the Classification Study	2-25
Training Panelists for the Booklet Classifications.....	2-26
The Practice Session of Booklet Classifications.....	2-26
The Booklet Classifications	2-27
Comparisons of Each Teacher's Three Classifications of Their Own Students	2-28
Panelists' Evaluations of the Procedures and Influences on Their Classifications	2-31
Responses to Questions Regarding Fundamental Features of the Classification Procedures	2-31
Identification of Factors That Influenced Panelists' Judgments	2-34
Influences Reported in Second Classification of Overall Civics Knowledge and Skills.....	2-36

Influence of Overall Knowledge and Skills in All Subjects.....	2-36
Influence of Overall Knowledge and Skills in Civics	2-37
Influence of Students' Grade in Teachers' Course.....	2-37
Influence of Items on the Civics NAEP	2-37
Influence of Length of Special NAEP and Motivation	2-37
Comments and Open-Ended Responses About Factors Influencing Classifications	2-37
Process Evaluation Questionnaires.....	2-37
Conclusions	2-41
Discussion of Findings	2-41
References	2-43

CIVICS ITEM CLASSIFICATION STUDY: 1998 CIVICS NAEP

Susan Cooper Loomis
ACT, Inc.

EXECUTIVE SUMMARY

ACT conducted this small item classification study at the request of the Achievement Levels Committee of the National Assessment Governing Board (NAGB). The purpose of the study was to determine whether there was evidence of a reasonable correspondence between the Civics NAEP Achievement Levels Descriptions (ALDs) and the performances of students within achievement levels. Is there evidence that students performing within the cutscore ranges *know and can do* the types of things that the ALDs require for performance within each level?

This type of research requires the use of a response probability to locate or map items onto the score scale. The choice of response probabilities will, in large part, determine the particular match of items to levels. NAGB has not established a response probability to use in the achievement levels-setting (ALS) process, however, so three different probabilities were examined.

- In general, the findings of the study showed that there was a reasonable correspondence between the performance of students in each achievement level category and items that represent the knowledge and skills that students *should* have, according to the ALDs.
- There was little difference by item type (i.e., multiple choice and constructed response items) in the correspondence of classifications based on teachers' judgments and student performance data.
- The highest rate of agreement in item classifications according to the ALDs was reached by twelfth grade teachers (98%) and the lowest rate was by eighth grade teachers (71%). Across the three grades, the average rate of agreement was 86%.
- Further evaluations were made of the impact of response probabilities on the correspondence between item classifications based on teachers' judgments relative to those based on response probabilities.
 - A 65% response probability showed the highest correspondence with teachers' classifications of items.
- Further evaluations were made of the impact of using a correction for guessing on the correspondence between item classifications based on teachers' judgments relative to those based on response probabilities, with and without a correction for guessing.
 - Using a correction for guessing on multiple choice items increased the correspondence with teachers' classifications of items. A 65% probability of correct response, corrected for guessing, is mapped as if the response probability were 74%.
 - A surprisingly large percentage of items in the 1998 Civics NAEP item pool had relatively high guessing parameters. Forty percent of the multiple choice items at grade 4, 54% at grade 8 and 69% at grade 12 had guessing parameters that exceeded the chance probability (25%) of randomly guessing the correct response.

- The relatively high potential for student guessing would likely lead to a lower correspondence between teachers' judgments of item difficulty and student performances.

CIVICS ITEM CLASSIFICATION STUDY: 1998 CIVICS NAEP

Susan Cooper Loomis¹

ACT, Inc.

BACKGROUND INFORMATION ABOUT THE STUDY

The Achievement Levels Committee (ALC) of the National Assessment Governing Board (NAGB) requested in March 1999 that ACT prepare a report on the relationship between items and achievement levels for the National Assessment of Educational Progress (NAEP) in Civics. ACT designed and implemented the Civics Item Classification Study (CICS) to provide classification data for reporting to the NAGB Achievement Levels Committee. The Committee wanted to have the data to review prior to making their decisions regarding setting the civics achievement levels

During a meeting in April 1998 to review the finalized achievement levels descriptions, the Achievement Levels Committee had asked ETS to conduct a similar study. ETS conducted an item classification study to determine the extent to which the revised achievement levels descriptions (ALDs) recommended by ACT for use in the achievement levels-setting (ALS) process were represented in the item pool for each grade. ETS conducted that study and found that there was generally a very high level of correspondence between the ALDs and item pools for each grade. There were some items related to reading maps, charts, and graphs for which there were no descriptors, and this review led to revisions in some ALDs to include those skills (Loomis & Hanick, 2000).

The request from the Achievement Levels Committee for the current study was not specific, but the general purpose of the study was related to the validity of the achievement levels. They wanted to know whether there was evidence of correspondence between the knowledge and skills that students performing in each achievement level category demonstrated and the knowledge and skills that the achievement levels descriptions required. ACT agreed to review the item classifications prepared by ETS in 1998 and to compute descriptive statistics for items classified in each achievement level category.

ACT Project Staff discussed this request with the Technical Advisory Team (TAT) in March². TAT was opposed to the study because it represents a perspective on standard setting that is contrary to that used by ACT for NAGB. Since TAT recognized that ACT needed to respond to the formal request by the Achievement Level Committee, however, they recommended the following strategy. They recommended that ACT first request more specific information from ETS regarding the methodology they used for classifying the items. The concern of TAT was that the classification lacked reliability. TAT strongly recommended against using the ETS classifications for further analyses unless there were evidence that the classifications were reliable. If the Project Staff were convinced that the classifications had followed systematic

¹ Dr. Wen-Ling Yang, Educational Testing Service, analyzed the data used in this report during her tenure at ACT. The author wishes to acknowledge her significant contribution to this research.

² There was no time to discuss this study request with the Technical Advisory Committee on Standard Setting (TACSS) since the report was needed by NAGB before their next meeting in early May 1999. Data analyses continued after the May 1999 decision by NAGB to adopt the achievement levels as recommended. TACSS reviewed the data before they were presented to NAGB in May. TACSS made recommendations for additional data analyses and modifications that were completed in October 1999.

procedures and were likely to represent a relatively reliable set of classifications, then those classifications could be used. If not, TAT recommended that ACT conduct a small classification study to accomplish this. TAT suggested that ACT recruit a small number of teachers from the Iowa City School District to participate in the study.

ACT discussed the 1998 study with ETS and confirmed that only one person had classified the items. The original purpose of the classification had not required that more persons in the classification process. The person who classified the items discussed his personal reservations about such a task and commented on his concerns about other uses of the data.

PANELISTS

ACT contacted the social studies curriculum director for the Iowa City School District. The curriculum director contacted the grade 12 social studies teachers who teach government/civics courses, as well as the social studies teachers in grade 8 and teachers in grade 4. Teachers were enlisted as panel members on a first-come-first-serve basis.

ACT offered to pay \$200 for the task that was estimated to take one full day. ACT offered a choice to the district coordinator between scheduling the study for several weekday afternoons/evenings versus all day Saturday. The advice was to conduct the study on Saturday, April 10.

Three panel members for each of the three grades tested in NAEP were recruited for the study. One grade 8 teacher had participated in the first field trial for civics conducted by ACT in February 1998. That study used geography data, however, and there was no reason to believe that the field trial experience would impact the substantive aspects of this study for this panelist.

The grade 4 panelists were all women. They were from three different elementary schools. Two men and one woman served on the grade 8 panel. There are only two public middle schools in the Iowa City School District, and both were represented on the grade 8 panel. The grade 12 panelists were all men. These three teachers were from the same school, although there are two public high schools in the district.

MATERIALS

ADVANCE MATERIALS

Panelists were contacted by email and sent a rather detailed description of the study. They were then mailed a set of advance materials and asked to review them before arriving for the study. These materials included NAGB policy definitions, the finalized achievement levels descriptions for all three grades, the Civics Framework, the NAEP Guide, and a brochure with information about NAGB. Copies of letters sent to panelists are included in Appendix A.

MATERIALS FOR THE STUDY

The item rating booklets used in the ALS process were used in this study. These booklets include all items, by block, in the grade level item pool and the scoring rubrics for each item. The item rating booklets are organized and color-coded according to the blocks included in the rating pool for each rating group. There were enough booklets left over to have all panelists use the same booklet form.

In addition, panelists were provided three student papers scored at each rubric score greater than 1 (inappropriate). These papers were for reference in the process of classification to give the teachers examples of the quality of response associated with each credited response.

Two scannable classification forms were designed for use by panelists. One form was for individual, independent classifications, and the second was for the grade level adjudicated classification. Panelists had a space to note “how sure” they were about their classification of each item. Please see Figures 1 and 2 for examples of the rating forms used in the study.

TRAINING FOR THE CLASSIFICATION TASK

The NAEP ALS Project Director provided a brief overview of the steps in the ALS process completed to date. She provided a brief overview of the process by which the achievement levels descriptions had been revised, and she reviewed the achievement levels descriptions with the panelists.

Panelists were instructed in the task. They were asked to classify each item at the lowest achievement level for which students were likely to give a correct response. This judgement was to be based on the knowledge and skill required to correctly respond to the question or task relative to the achievement levels descriptions for the grade. Polytomously-scored items were considered as multiple items, one item for each score of 2 or higher. That is, each credited response code was considered a separate item for classification purposes.

Panelists were asked to classify each item at each of the three achievement levels, based on the descriptions. Items that appeared to require a lower level of knowledge or skill for correct response were to be classified as *below Basic*. Items that were judged to be very difficult were to be classified as Advanced, even if the items seemed to require a much higher level of knowledge and skill than the Advanced ALD.

Panelists were also asked to mark how sure they were in their classification for each item. Their choices were *very sure*, *fairly sure*, and *not sure*.

Panelists were instructed that they would need to pace themselves carefully. They were urged to record their first judgement. They would need to classify items at an average rate of one item every 1.5 minutes in order to finish the task in the time allotted. This estimate would provide enough time for an independent round of classification, with approximately two hours left for a group round.

THE CLASSIFICATION PROCESS

Panelists began classifying items at 9:15 A.M. After 30 minutes, panelists were given a time check and told that this was approximately the time allotted for a block of items. All panelists had completed one block in that amount of time.

Panelists completed the task much more rapidly than anticipated. The first panelist to complete the independent classification task was a fourth grade teacher. Two of the fourth grade teachers had completed their independent ratings before 11:00 A.M.

After completing the independent round, panelists were given another form for recording their group classification of each item. The group form was the same as the individual form, but the group did not record an evaluation regarding their certainty for each item classification.

Panelists were instructed to read through their classifications for each item and pause to discuss only those items for which there was disagreement. The following rules were read to the panelists before they began their classifications, and they were told that these rules were in effect for their group classifications.

- a) Consensus is the goal, but it is not a requirement. If the three panelists cannot agree, then each should record their own judgment. The classification of the majority will be used. If none agree, then the item cannot be classified, and that will be reported.
- b) If two panelists have completed their classifications ahead of the likely completion of the third panelist (e.g., as much as 20 minutes ahead), those two should begin the adjudication process.
- c) If none of the panelists at a grade are finished in time for a complete review of the items—or if the adjudication process takes longer than anticipated—the back-up strategy is for panelists discuss *only* those items for which all three disagreed.

The two fourth grade panelists who completed their independent classifications early began discussing their ratings before the third panelist had completed her independent classifications. Panelists for the eighth grade and twelfth grade waited until all three were ready to begin the group classification. All panelists had completed their independent classifications by noon. A twelfth grade panelist was the last to complete his independent classifications.

Interestingly, the grade 12 panelists completed their group adjudication process first, followed by the grade 8 panelists. The grade 4 group started the adjudication process about an hour before the other two grade groups, and they completed the task at least 30 minutes after the grade 8 panelists. All of the grade 12 panelists were teachers from the same high school; two of the three grade 8 panelists were from the same school; and none of the grade 4 teachers were from the same school.

DATA ANALYSIS AND REPORTING

AGREEMENT AMONG TEACHERS

Percentages of items for which all three teachers in each grade group agreed were computed. Table 1 shows the overall percentages of agreement for each grade for all items at the grade level and for multiple choice and constructed response items separately. There was no pattern of higher or lower agreement with respect to item type, i.e. multiple choice and constructed response items.

Grade 12 teachers had the highest agreement. They reached agreement on the classification of 98% of the items. Of the 184 items classified, they disagreed on only three items. These three items were in the same item block, and all were dichotomous items. One teacher classified these three multiple-choice items below the Basic level and the other two teachers classified them at the Basic level. All three grade 12 panel members classified the other 181 items at the same level.

Grade 8 teachers had the lowest rate of agreement in that they reached agreement on only 71% of the items. The proportion of agreement was about the same for both multiple choice and constructed response items. Grade 8 teachers reached unanimous agreement on 130 of the 182

items in the grade 8 item pool. There was a definite pattern of agreement between the same two panelists and disagreement by the third member. One grade 8 teacher disagreed with the other two on about 25% of the items in the pool. The disagreement did not reflect a systematic pattern of higher or lower classification by the dissenting panelist.

Grade 4 teachers reached agreement on 88% of the items. There were 113 items in the grade 4 item pool, and the grade four teachers agreed unanimously on 99 of those items. The disagreement was again patterned. One specific teacher tended to disagree with the other two. She disagreed on about 11% of the total items. Again, the disagreement did not reflect a consistently higher or lower judgement by the dissenting panelist. Grade 4 panelists did have a higher agreement on multiple choice items than on constructed response items.

ITEM DIFFICULTY AND CLASSIFICATION LEVEL

Classifications were used to compute descriptive statistics for items at each level at each grade. Data were analyzed to determine the average probability of correct response for items classified at each achievement level for each grade. Those results are reported in Table 2. This analysis is rather straightforward. The data used here for multiple choice items are the percentage of students correctly answering each item. For constructed response items, the percentage of students giving a correct response was operationalized as the percentage of students earning at least partial credit, i.e. the percent scoring 2, 3, or 4 on constructed response items. The average percentages are computed for all items classified at each level. The data in Table 2 show that panelists were able to judge items and to match them to achievement levels categories so that relatively more difficult items were judged to require relatively higher levels of student achievement and relatively easier items were judged to require relatively lower levels of student achievement. The exception to this is for the few items classified as *Below Basic*. Some rather difficult items were classified in the Below Basic category. Note that judgments by fourth grade teachers classified no items at the Below Basic level.

Tables 3-5 report data for the average conditional probabilities of items classified at each level. The data in those tables are *conditional* data, meaning that they are based on performances of students scoring within each achievement level range. The average conditional probability of correct response was first computed for each item. This average conditional probability is the average probability of correct response for student scores within a specific achievement level category. The average reported is the average for the items classified at a specific level. For example, if 15 items were classified as meeting the requirements of Basic level performance, the average conditional probability of correct response for the 15 items was computed for the Basic score range. It was also computed for the below Basic range, the Proficient range, and the Advanced range. The conditional descriptive statistics are reported across each level for each set of items classified at each level. Items classified at the Basic level should have a higher probability of correct response at the Proficient and Advanced levels if the achievement levels descriptions have served well to guide the classifications. In all cases, the expected pattern is observed.

The range of average conditional probabilities presents an interesting pattern. The average conditional probability at the Basic level for items classified as *Basic* is lower than the average conditional probability at the Proficient level for items classified as *Proficient* and both are lower than the average conditional probability at the Advanced level for items classified as *Advanced*. The average conditional probabilities range from the upper 50's for the Basic range of scores for items classified at the Basic level to the low or middle 70's for the Advanced range of scores for items classified at the Advanced level. This means that items classified at the Basic level were

relatively more difficult for students performing at that level than items classified at the Advanced level for students performing at that level. Perhaps the teachers required a relatively higher level of mastery over material for performances at higher levels of achievement.

ANALYSES USING DIFFERENT RESPONSE PROBABILITY CRITERIA

The correspondence between judgments by teachers and classifications of items based on empirical performance data was examined in detail. In order to analyze the agreement between teachers' judgments and classifications based on performance data, it was necessary to select a probability of correct response to use for mapping items to the score scale. The mapping allows one to classify items into each achievement level category, based on the score associated with the items. Response probabilities typically used in the NAEP program are .50, .65, and .80. (See Zwick, Senturk, Wang & Loomis, 2000). Data reported in Tables 6-8 present the correspondence between judgmental and empirical classifications for a response probability of .50. Tables 9-11 show the correspondence for classifications using a response probability of .65, and Tables 12-14 show the correspondence for a response probability of .80. The number of items included in the item count again includes one *item* for each credited response category.

In general, the highest correspondences between judgmental and empirical classifications were found for the classifications based on a response probability of .65.

ANALYSIS OF THE EFFECT OF USING A CORRECTION FOR GUESSING

The three-parameter IRT model used for scaling NAEP data includes an estimate of the effect of guessing for multiple choice items. A *correction for guessing* is typically used for NAEP reports that include multiple choice item maps. The reason for using a correction for guessing is to make the relative difficulty of multiple choice and constructed response items more equal. Since random guessing can result in the correct response, the probability of correct response should be higher for multiple choice items than for constructed responses, *ceteris paribus*. Rather than using the actual item parameter estimate (the *c* parameter), a constant correction value is used for mapping multiple choice items. A probability of .74 is the probability value when a response probability of .65 is *corrected for guessing*, for example. This is based on a .25 probability of randomly guessing the correct response for multiple choice items having four choices, such as in the Civics NAEP. Values of .625 and .85 are the *corrected* values for response probabilities of .50 and .80, respectively.

The correspondence between teachers' judgments and the empirical classifications were compared for each response probability with and without a correction for guessing. Results of those analyses are summarized in Tables 15 and 16. Agreement between the two classifications is reported as a *Hit*. If the comparison shows that teachers judged the item at a higher level than the empirical classification level, the classification is labeled an *overestimate*. If the teachers judged the item at a lower level than the empirical classification level, the classification is labeled an *underestimate*. The results in Table 17 provide a summary of the comparisons. In Table 17, the differences between the classification data in Table 16 (based on no correction for guessing) and those in Table 15 (based on a correction for guessing) are reported. The results in Table 17 indicate that the correction for guessing tended to be associated with more underestimation, as would be expected.

ANALYSIS OF IRT PSEUDO-GUESSING PARAMETERS FOR CIVICS ITEMS

The discussions with technical advisors led to more interest in the question of whether a correction for guessing should be used in mapping the items for analysis. Members of the Technical Advisory Committee on Standard Setting had recommended the analyses be based on empirical performance probabilities with *no correction* for guessing. They reasoned that panelists are unlikely to factor in guessing explicitly when making their classification judgments. Further, they reasoned that the constant correction factor did not accurately represent the estimates of the effects of guessing. The analyses of the magnitude of the empirical estimates of the guessing factor are presented in Table 18.

Perhaps the correspondence between teacher judgments and the empirical classifications was impacted by the incidence of guessing. The data presented in Table 18 suggest that the incidence of guessing was quite high for the Civics NAEP. Over two-thirds of the multiple choice items at grade 12 had an estimated guessing parameter in excess of .25 (the random guessing probability, given four choices on the multiple choice items). The high probability of correct responses through guessing would likely diminish the correspondence between the judgements of the teacher panelists in the study and the empirical classifications.

SUMMARY

The correspondence between teacher judgements of the match of items to achievement levels descriptions and student performance on the items, relative to the achievement levels cutscores was judged to be reasonably strong. This was taken as evidence of the validity of the achievement levels-setting process. Teachers used the achievement levels descriptions to classify items. Those classifications generally corresponded to the classifications based on performances of students who scored in the range of the achievement levels cutscores. There was little difference between multiple choice and constructed response items in the correspondence of classifications based on teachers' judgments and empirical performance data. Agreement among the teachers was highest for grade 12 and lowest for grade 8.

The analyses of these data led to further findings of interest. For example, the classifications by these teachers match an empirical classification based on a 65% probability of correct response better than either a 50% or 80% probability. The highest rate of correspondence was found using a response probability of 65% and a correction for guessing. A surprisingly large percentage of items in the 1998 Civics NAEP item pool had relatively high *guessing* parameters. This means that students in low performance ranges had a *higher-than-random-chance* probability of a correct response on a relatively large number of items. These items clearly provide little information for assessing students in the lower ranges of performance. The relatively high potential for guessing would likely lead to a lower correspondence between classifications based on teachers' judgments relative to the achievement levels descriptions of how students should perform and classifications based on student performance on items.

REFERENCE

- Loomis, S.C. & Hanick, P.L. (2000). *Setting standards for the 1998 NAEP in civics and writing: Finalizing the achievement levels descriptions*. Iowa City, IA: ACT.
- Zwick, R., Senturk, D., Wang, J., & Loomis, S.C. (2000). *An Investigation of Alternative Methods for Item Mapping in the National Assessment of Educational Progress*. Iowa City, IA: ACT, Inc.

Table 1
Percent of Agreement by Teachers on Classification of 1998 NAEP Civics Items

Grade	Item Type	Number of Items with Perfect Agreement	Total Number of Items	Percent of Agreement
4	Dichotomous	61	68	90
	Polytomous	38	45	84
	Both	99	113	88
8	Dichotomous	87	122	71
	Polytomous	43	60	72
	Both	130	182	71
12	Dichotomous	120	123	98
	Polytomous	61	61	100
	Both	181	184	98

Table 2
Average Percent Correct for Items Classified in Each Achievement Level Category

Classification Level	Number of Items	Average % Correct	SD	Minimum	Maximum
Grade 4					
Below Basic	0	-	-	-	-
Basic	37	52.3	24.8	8.4	93.0
Proficient	61	39.4	20.7	1.7	77.8
Advanced	15	36.7	23.8	2.5	73.4
Grade 8					
Below Basic	2	35.6	26.3	17	54.2
Basic	49	53.6	19.8	15.5	89.4
Proficient	110	41.2	20.2	4.4	84.5
Advanced	21	30.2	21.7	2	76.8
Grade 12					
Below Basic	2	28.6	0.4	28.3	28.9
Basic	82	53.9	20.3	8.6	90.0
Proficient	88	41.7	41.7	5.8	83.1
Advanced	12	31.1	31.1	3.1	53.8

Table 3
Descriptive Statistics for Items Classified
in Each Achievement Level Category
Grade 4

Level	Avg Cond Probability	SD	Minimum Avg Cond P	Maximum Avg Cond P	Median Avg Cond P
Basic=37 items					
Below Basic	43.4	21.7	4.7	85.8	42.6
Basic	66.2	21.2	15.4	95.9	69.6
Proficient	84.1	14.5	38.3	98.7	89.3
Advanced	93.9	7.5	72.5	99.7	97.2
Proficient=61 items					
Below Basic	23.8	15.2	0.3	54.4	24.5
Basic	40.8	22.5	1.4	84.5	39.6
Proficient	62.0	24.3	3.9	98.3	66.1
Advanced	80.6	19.4	10.4	99.8	86.0
Advanced=15 items					
Below Basic	22.9	17.0	0.1	54.8	28.0
Basic	36.1	25.1	1.6	76.3	36.0
Proficient	53.9	29.7	7.7	92.4	53.5
Advanced	71.7	25.9	25.0	98.8	81.2

Table 4
Descriptive Statistics for Items Classified
in Each Achievement Level Category
Grade 8

Level	Avg Cond Probability	SD	Minimum Avg Cond P	Maximum Avg Cond P	Median Avg Cond P
Below Basic=2 items					
Below Basic	31.8	34.6	7.3	56.3	31.8
Basic	45.5	40.9	16.6	74.4	45.5
Proficient	56.9	40.5	28.2	85.5	56.9
Advanced	67.4	35.4	42.3	92.4	67.4
Basic=49 items					
Below Basic	38.3	17.8	1.5	80.4	35.2
Basic	57.3	21.7	16.6	95.1	60.3
Proficient	77.6	17.6	24.2	99.2	80.6
Advanced	91.2	9.5	54.0	99.9	94.8
Proficient=110 items					
Below Basic	25.9	16.1	0.0	74.3	26.9
Basic	43.4	22.3	2.0	90.4	43.2
Proficient	67.7	22.0	9.7	98.9	72.3
Advanced	86.6	15.5	15.5	99.9	92.5
Advanced=21 items					
Below Basic	18.0	18.3	0.0	49.3	8.8
Basic	32.2	26.2	0.7	82.3	23.7
Proficient	52.0	26.6	6.8	96.1	46.5
Advanced	74.9	18.3	31.4	99.2	72.8

Table 5
Descriptive Statistics for Items Classified
in Each Achievement Level Category
Grade 12

Level	Avg Cond Probability	SD	Minimum Avg Cond P	Maximum Avg Cond P	Median Avg Cond P
Below Basic=2 items					
Below Basic	43.0	30.3	21.5	64.4	43.0
Basic	70.7	25.6	52.6	88.8	70.7
Proficient	88.7	11.0	80.9	96.4	88.7
Advanced	97.6	2.0	96.2	99.0	97.6
Basic=82 items					
Below Basic	40.0	18.6	0.4	90.8	38.2
Basic	61.5	20.0	4.9	97.8	63.4
Proficient	81.3	15.0	21.0	99.4	86.1
Advanced	93.8	7.4	54.8	99.9	96.2
Proficient=88 items					
Below Basic	25.7	15.8	0.0	71.6	28.4
Basic	42.5	19.9	1.5	89.7	42.7
Proficient	65.4	20.1	15.0	97.7	67.7
Advanced	85.2	14.3	39.8	99.6	88.8
Advanced=12 items					
Below Basic	18.9	14.3	0.1	37.5	22.9
Basic	28.1	19.5	1.1	53.7	31.4
Proficient	47.7	24.9	7.8	76.5	52.7
Advanced	74.2	22.3	34.7	94.3	86.1

Table 6
Comparisons of Item Classification Outcomes for Grade 4:
Teacher Judgments v Performance Level at RP .50

		Teachers' Classification			
		Basic	Proficient	Advanced	Total
Empirical Classification RP=0.50	Below Basic				
	n=	24	12	2	38
	Table %	21.24	10.62	1.77	33.63%
	Row %	63.16	31.58	5.26	
	Column %	64.86	19.67	13.33	
	Basic				
	n=	9	23	5	37
	Table %	7.96	20.35	4.42	32.74%
	Row %	24.32	62.16	13.51	
	Column %	24.32	37.70	33.33	
	Proficient				
	n=	4	16	4	24
	Table %	3.54	14.16	3.54	21.24%
Row %	16.67	66.67	16.67		
Column %	10.81	26.23	26.67		
Advanced					
n=	0	10	4	14	
Table %	0.00	8.85	3.54	12.39%	
Row %	0.00	71.43	28.57		
Column %	0.00	16.39	26.67		
Total					
	37	61	15	113	
	32.74%	53.98%	13.27%	100.00%	

Table 7
Comparisons of Item Classification Outcomes for Grade 8:
Teacher Judgments v Performance Level at RP .50

		Teachers' Classification				
		Below Basic	Basic	Proficient	Advanced	Total
Empirical Classification RP=0.50	Below Basic					
	n=	1	16	15	4	36
	Table %	0.55	8.79	8.24	2.20	19.78%
	Row %	2.78	44.44	41.67	11.11	
	Column %	50.00	32.65	13.64	19.05	
	Basic					
	n=	0	21	52	4	77
	Table %	0.00	11.54	28.57	2.20	42.31%
	Row %	0.00	27.27	67.53	5.19	
	Column %	0.00	42.86	47.27	19.05	
	Proficient					
	n=	0	11	35	6	52
	Table %	0.00	6.04	19.23	3.30	28.57%
Row %	0.00	21.15	67.31	11.54		
Column %	0.00	22.45	31.82	28.57		
Advanced						
n=	1	1	8	7	17	
Table %	0.55	0.55	4.40	3.85	9.34%	
Row %	5.88	5.88	47.06	41.18		
Column %	50.00	2.04	7.27	33.33		
Total						
	2	49	110	21	182	
	1.10%	26.92%	60.44%	11.54%	100.00%	

Table 8
Comparisons of Item Classification Outcomes for Grade 12:
Teacher Judgments v Performance Level at RP .50

	Teachers' Classification				
	Below Basic	Basic	Proficient	Advanced	Total
Below Basic					
n=	1	40	12	0	53
Table %	0.54	21.74	6.52	0.00	28.80%
Row %	1.89	75.47	22.64	0.00	
Column %	50.00	48.78	13.64	0.00	
Basic					
n=	1	31	42	4	78
Table %	0.54	16.85	22.83	2.17	42.39%
Row %	1.28	39.74	53.85	5.13	
Column %	50.00	37.80	47.73	33.33	
Proficient					
n=	0	10	25	4	39
Table %	0.00	5.43	13.59	2.17	21.20%
Row %	0.00	25.64	64.10	10.26	
Column %	0.00	12.20	28.41	33.33	
Advanced					
n=	0	1	9	4	14
Table %	0.00	0.54	4.89	2.14	7.61%
Row %	0.00	7.14	64.29	28.57	
Column %	0.00	1.22	10.23	33.33	
Total					
	2	82	88	12	184
	1.09%	44.57%	47.83%	6.52%	100.00%

Table 9
Comparisons of Item Classification Outcomes for Grade 4:
Teacher Judgments v Performance Level at RP .65

		Teachers' Classification			
		Basic	Proficient	Advanced	Total
Empirical Classification RP=0.65	Below Basic				
	n=	12	3	1	16
	Table %	10.62	2.65	0.88	14.16%
	Row %	75.00	18.75	6.25	
	Column %	32.43	4.92	6.67	
	Basic				
	n=	16	17	3	36
	Table %	14.16	15.04	2.65	31.86%
	Row %	44.44	47.22	8.33	
	Column %	43.24	27.87	20.00	
	Proficient				
	n=	8	20	4	32
	Table %	7.08	17.70	3.54	28.32%
	Row %	25.00	62.50	12.50	
	Column %	21.62	32.79	26.67	
	Advanced				
n=	1	21	7	29	
Table %	0.88	18.58	6.19	25.66%	
Row %	3.45	72.41	24.14		
Column %	2.70	34.43	46.67		
Total					
	37	61	15	113	
	32.74%	53.98%	13.27%	100.00%	

Table 10
Comparisons of Item Classification Outcomes for Grade 8:
Teacher Judgments v Performance Level at RP .65

	Teachers' Classification				
	Below Basic	Basic	Proficient	Advanced	Total
Below Basic					
n=	1	10	7	1	19
Table %	0.55	5.49	3.85	0.55	10.44%
Row %	5.26	52.63	36.84	5.26	
Column %	50.00	20.41	6.36	4.76	
Basic					
n=	0	21	35	5	61
Table %	0.00	11.54	19.23	2.75	33.52%
Row %	0.00	34.43	57.38	8.20	
Column %	0.00	42.86	31.82	23.81	
Proficient					
n=	0	15	47	3	65
Table %	0.00	8.24	25.82	1.65	35.71%
Row %	0.00	23.08	72.31	4.62	
Column %	0.00	30.61	42.73	14.29	
Advanced					
n=	1	3	21	12	37
Table %	0.55	1.65	11.54	6.59	20.33%
Row %	2.70	8.11	56.76	32.43	
Column %	50.00	6.12	19.09	57.14	
Total	2	49	110	21	182
	1.10%	26.92%	60.44%	11.54%	100.00%

Table 11
Comparisons of Item Classification Outcomes for Grade 12:
Teacher Judgments v Performance Level at RP .65

		Teachers' Classification				
		Below Basic	Basic	Proficient	Advanced	Total
Empirical Classification RP=0.65	Below Basic					
	n=	1	20	3	0	24
	Table %	0.54	10.87	1.63	0.00	13.04%
	Row %	4.17	83.33	12.50	0.00	
	Column %	50.00	24.39	3.41	0.00	
	Basic					
	n=	1	41	23	0	65
	Table %	0.54	22.28	12.50	0.00	35.33%
	Row %	1.54	63.08	35.38	0.00	
	Column %	50.00	50.00	26.14	0.00	
	Proficient					
	n=	0	18	44	8	70
	Table %	0.00	9.78	23.91	4.35	38.04%
	Row %	0.00	25.71	62.86	11.43	
	Column %	0.00	21.95	50.00	66.67	
	Advanced					
n=	0	3	18	4	25	
Table %	0.00	1.63	9.78	2.17	13.59%	
Row %	0.00	12.00	72.00	16.00		
Column %	0.00	3.66	20.45	33.33		
Total						
	2	82	88	12	184	
	1.09%	44.57%	47.83%	6.52%	100.00%	

Table 12
Comparisons of Item Classification Outcomes for Grade 4:
Teacher Judgments v Performance Level at RP .80

		Teachers' Classification			
		Basic	Proficient	Advanced	Total
Empirical Classification RP=0.80	Below Basic				
	n=	6	0	0	6
	Table %	5.31	0.00	0.00	5.31%
	Row %	100.00	0.00	0.00	
	Column %	16.22	0.00	0.00	
	Basic				
	n=	14	12	3	29
	Table %	12.39	10.62	2.65	25.66%
	Row %	48.28	41.38	10.34	
	Column %	37.84	19.67	20.00	
	Proficient				
	n=	11	18	3	32
	Table %	9.73	15.93	2.65	28.32%
	Row %	34.38	56.25	9.38	
	Column %	29.73	29.51	20.00	
	Advanced				
n=	6	31	9	46	
Table %	5.31	27.43	7.96	40.71%	
Row %	13.04	67.39	19.57		
Column %	16.22	50.82	60.00		
Total					
	37	61	15	113	
	32.74%	53.98%	13.27%	100.00%	

Table 13
Comparisons of Item Classification Outcomes for Grade 8:
Teacher Judgments v Performance Level at RP .80

		Teachers' Classification				
		Below Basic	Basic	Proficient	Advanced	Total
Empirical Classification RP=0.80	Below Basic					
	n=	0	4	1	0	5
	Table %	0.00	2.20	0.55	0.00	2.75%
	Row %	0.00	80.00	20.00	0.00	
	Column %	0.00	8.16	0.91	0.00	
	Basic					
	n=	1	15	20	3	39
	Table %	0.55	8.24	10.99	1.65	21.43%
	Row %	2.56	38.46	51.28	7.69	
	Column %	50.00	30.61	18.18	14.29	
	Proficient					
	n=	0	16	47	4	67
	Table %	0.00	8.79	25.82	2.20	36.81%
	Row %	0.00	23.88	70.15	5.97	
	Column %	0.00	32.65	42.73	19.05	
	Advanced					
n=	1	14	42	14	71	
Table %	0.55	7.69	23.08	7.69	39.01%	
Row %	1.41	19.72	59.15	19.72		
Column %	50.00	28.57	38.18	66.67		
Total						
	2	49	110	21	182	
	1.10%	26.92%	60.44%	11.54%	100.00%	

Table 14
Comparisons of Item Classification Outcomes for Grade 12:
Teacher Judgments v Performance Level at RP .80

		Teachers' Classification				
		Below Basic	Basic	Proficient	Advanced	Total
Empirical Classification RP=0.80	Below Basic					
	n=	1	10	1	0	12
	Table %	0.54	5.43	0.54	0.00	6.52%
	Row %	8.33	83.33	8.33	0.00	
	Column %	50.00	12.20	1.14	0.00	
	Basic					
	n=	0	26	7	0	33
	Table %	0.00	14.13	3.80	0.00	17.93%
	Row %	0.00	78.79	21.21	0.00	
	Column %	0.00	31.71	7.95	0.00	
	Proficient					
	n=	1	32	34	4	71
	Table %	0.54	17.39	18.48	2.17	38.59%
	Row %	1.41	45.07	47.89	5.63	
	Column %	50.00	39.02	38.64	33.33	
	Advanced					
n=	0	14	46	8	68	
Table %	0.00	7.61	25.00	4.35	36.96%	
Row %	0.00	20.59	67.65	11.76		
Column %	0.00	17.07	52.27	66.67		
Total						
	2	82	88	12	184	
	1.09%	44.57%	47.83%	6.52%	100.00%	

Table 15
Percent Agreement Between Teachers' Classifications of Civics Items
Using Different Response Probabilities and a Correction for Guessing on Multiple-Choice Items

Agreement/ Disagreement	Response Probability	Percent Agreement by Grade Level		
		4	8	12
Hit	0.65	39%	45%	52%
	0.50	42	42	46
	0.80	35	41	34
Overestimation	0.65	29	26	20
	0.50	43	42	39
	0.80	17	13	07
Underestimation	0.65	31	29	29
	0.50	15	15	16
	0.80	49	47	59

Table 16
Percent Agreement Between Teachers' Classifications of Civics Items
Using Different Response Probabilities and No Correction for Guessing on Multiple-Choice Items

Agreement/ Disagreement	Response Probability	Percent Agreement by Grade Level		
		4	8	12
Hit	0.65	38%	45%	49%
	0.50	26	35	33
	0.80	36	42	38
Overestimation	0.65	35	34	29
	0.50	62	53	55
	0.80	21	18	12
Underestimation	0.65	27	22	22
	0.50	12	12	11
	0.80	42	41	51

Table 17
Rates of Agreement for Teachers' Classifications of Civics Items v. Empirical Classifications:
Difference Between Rates Without the Response Probability Adjusted for Guessing and
With Response Probability Adjusted for Guessing

Agreement/ Disagreement	Response Probability	Grade Level		
		4	8	12
Hit	0.65	-0.01	-0.01	-0.03
	0.50	-0.16	-0.07	-0.13
	0.80	0.02	0.01	0.03
Overestimation	0.65	0.06	0.07	0.10
	0.50	0.19	0.11	0.17
	0.80	0.04	0.05	0.05
Underestimation	0.65	-0.04	-0.07	-0.07
	0.50	-0.03	-0.04	-0.04
	0.80	-0.06	-0.06	-0.08

Note: Three response probabilities were used in this analysis: 0.65, 0.50, and 0.80. When the response probabilities for multiple-choice items were corrected for guessing, the probabilities were 0.74, 0.625, and 0.85, respectively. The differences reported here are difference for cells in Table 16 – Table 15.

Table 18
Percentages of Multiple-Choice Items for the 1998 Civics NAEP for which the
Pseudo-Chance Parameter (c) was Estimated to be Greater than the Probability Associated with
Random Guessing

Grade	Total # of MC Items	# of MC Items for which $c \geq .25$	Percentage
4	67	27	40.3%
8	121	65	53.7%
12	122	84	68.9%

Figure 1

Civics Item Classification Study

April 10, 1999

Individual Classification Form

Grade 8




ID

O
O
O

Name: _____

Block U2C5	Below Basic	Basic	Proficient	Advanced	Very Sure	Fairly Sure	Not Sure
1.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3. >=2	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
>=3	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
4.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
5.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
6. >=2	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
>=3	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
7.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
8.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
9.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
10. >=2	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
>=3	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
11.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
12.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
13. >=2	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
>=3	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
14.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
15.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
16.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
17.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
18.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
19.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 2

<p>Civics Item Classification Study April 10, 1999</p> <p>Group Classification Form</p> <p>Grade 4</p>	
---	---

ID

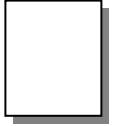
-
-
-

Name: _____

<i>Block U1C4</i>	Below Basic	Basic	Proficient	Advanced
1.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
4.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
5. >=2	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
>=3	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
6.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
7.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
8.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
9. >=2	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
>=3	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
10.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
11. >=2	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
>=3	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
12.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
13.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
14.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
15.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Appendix

Letters to Panelists



THE 1998 CIVICS NAEP VALIDATION STUDY OF ACHIEVEMENT LEVELS

EXECUTIVE SUMMARY

ACT has implemented holistic methods in studies for NAEP in five subjects: geography, US history, science, civics, and writing. Three different classification procedures have been implemented, and the results of these classifications have been compared to empirical score classifications based on NAEP achievement levels cutscores. The results of the Civics research confirmed the findings of the previous studies. By combining the different types of classifications into a single study performed by the same panelists, the results were more conclusive.

1. Teachers judge the overall knowledge and skills of their students to represent a level of achievement that is as high or higher than the classifications of the students' performance scores based on achievement level cutscores.
2. Teachers expect the performance of their students on the NAEP to be at the same or at a higher level than the empirical classifications of the students' performance scores based on achievement levels cutscores.
3. Panelists judge the performance of students on NAEP, represented in NAEP test booklets, to be at the same or at a lower level than the empirical classifications of the students' performance scores based on achievement level cutscores.
4. Of these three indicators of student achievement, cutscores will be set lowest if overall knowledge and skills of students is the dimension on which students are classified relative to achievement levels descriptions.
5. Of these three indicators of student achievement, cutscores will be set highest if student test booklets are used to represent student performance for classifications relative to achievement levels descriptions.
6. The ability to use NAEP achievement levels descriptions as the criterion for judging the achievement, performance, or expected performance of students is as high or higher than the ability of teachers to judge student scores on other tests.
7. Panelists will use a noncompensatory method for judging performance when the procedure is holistic.
8. Panelists will not approach a task in a holistic manner if they are given enough time to score individual items.
9. The ACT/NAGB ALS process results in cutscores that are higher than teachers' judgments of their own students' overall knowledge and skills and somewhat higher than teachers' judgments of their students' likely performance on the NAEP. The ACT/NAGB ALS process results in cutscores that are lower than those resulting from classifications of student test booklets.

THE 1998 CIVICS NAEP VALIDATION STUDY OF ACHIEVEMENT LEVELS

Susan Cooper Loomis
ACT, Inc.

with Patricia L. Hanick, Paul Nichols, Jim Sconing, Teri Fisher, ACT, Inc.
and Wen-Ling Yang, ETS

BACKGROUND OF THE STUDY

In 1995 ACT designed and conducted two validation studies related to the achievement levels-setting (ALS) process for geography and U.S. history (ACT, 1995a). The National Academy of Education (NAE) included the ALS process as part of the evaluation of the NAEP Trial State Assessments (NAE, 1993). The report of the NAE suggested that the achievement levels set for NAEP were fundamentally flawed. They argued in their report, and in subsequent papers and discussions (Shepard, 1995) that item-by-item rating methods will result in flawed standards. In particular, they argued that the standards would be higher than those set by holistic methods. They recommended that holistic methods be used for setting NAEP achievement levels.

As the NAGB ALS contractors, ACT designed several procedures to incorporate into the 1994 ALS process for geography and US history. These procedures were tested in pilot studies for the two subjects (ACT, 1995a; 1995b). Two related types of feedback were added as standard aspects of the NAEP ALS process as a result of these studies. Both the wholebooklet feedback and wholebooklet exercise are a part of the ACT/NAGB standard setting method. In addition, ACT designed validation research studies to examine the 1994 NAEP achievement levels for geography and US history. One study was called the Similarities Classification Study (SCS). Another study was called the Booklet Classification Study (BCS). The Booklet Classification Study design was implemented again for the 1996 Science NAEP, and it was implemented as a standard setting methodology in a field trial for the 1998 Writing NAEP.

The logic of the Similarities Classification Study (SCS) is to test whether teachers who participated in the ALS process are able to apply the achievement levels descriptions (ALDs) in a way that is consistent with their use of the descriptions in setting the achievement levels. In particular, teachers were asked to make judgments about the knowledge and skills of their own students and the performance of their own students relative to the achievement levels descriptions. Those judgments were to classify the knowledge and skills of each student with respect to the achievement levels descriptions and to classify the expected performance of each student on a special form of the NAEP with respect to the achievement levels descriptions.

These teachers are well trained in the ALDs during the ALS process. As a result of that experience, they should have a very good understanding of the meaning of the ALDs. If these teachers could use the ALDs to estimate achievement and performances of their students in a way consistent with the students' performance, then this would add support to the levels set. If not, then it seems unlikely that people who are not well trained in the ALDs and NAEP matters would be able to make reasonable interpretations about the meaning of the achievement levels. The teachers who have been ALS panel members provide the best case scenario. If they are unable to use the ALDs consistently, then one must assume that most people will misinterpret the outcomes of the NAEP ALS process.

The Booklet Classification Study (BCS) design has been implemented a number of times by ACT. In previous studies, the booklets have been NAEP booklets of students who were assessed for the regular assessment period of 50 minutes. The number of items included in those NAEP booklets was not sufficient to provide a reliable score to represent performance of an individual student. The composition and training of panelists in the BCS were as similar as possible to those in the achievement levels-setting process for NAEP. The task was holistic and required panelists to consider the overall performance of the student rather than to estimate the performance of students on each item. The booklet scores were not revealed to panelists, nor were scores on individual items within the booklets. Panelists classified booklets into achievement level categories using the achievement levels descriptions to judge the performance represented by the booklet as a whole.

NAGB requested that ACT implement similar studies as a part of the 1998 NAEP ALS validation research. The SCS and BCS designs were incorporated into a single study implemented for the 1998 Civics NAEP. The decision was made to conduct the study for civics only because of the high cost of the study and because the pattern of findings was the central interest—not the findings with respect to a specific subject or set of achievement levels. This combination of the two studies provided the opportunity to address some questions that were raised in the 1995 studies. In particular, the findings from the Similarities Classification Study showed that teachers classified their own students at the same level or one achievement level higher than the student's actual or empirical performance level on NAEP. That finding would indicate that the achievement levels had been set too high. A contrary result was found for the Booklet Classification Study. Panelists classified the performance level represented by booklets as being at the same or one level lower than the empirical score level classification relative to the cutscores. This finding would indicate that the achievement levels were set too low.

OVERVIEW OF THE CURRENT STUDY

Findings from the 1995 SCS indicated that the achievement levels were perhaps set too high. That is, teachers in both geography and US history tended to classify the knowledge and skills of their students at a level higher than the empirical performance level (i.e., higher than the achievement level category of their score on the special form of NAEP), and the same was true for the level at which they classified the expected performance of their students on the special form of NAEP. Findings from the 1995 SCS were countered by findings from the 1995 and 1997 BCS. The Booklet Classification Study for geography, US history, and science indicated that the achievement levels were perhaps set too low. That is, panelists, including teachers, nonteacher educators, and representatives of the general public for each subject, generally classified performance represented in student NAEP booklets at one achievement level lower than the empirical performance level indicated by the score for the booklet.

Whether that was a general finding related to an inherent difference in the judgment tasks or a difference due to other factors was not known. For example, we were unable to determine whether the two sets of panelists differed to an extent that would have accounted for the differences in the classifications for the two types of tasks. A further complication with the BCS design was that the booklets used were actual NAEP booklets selected on the basis of plausible values. NAEP does not include enough items to produce a reliable individual student score. ACT was somewhat skeptical about using these booklets to represent individual student performances. Raw scores on the NAEP booklets tend to correspond to lower levels of achievement than is the case for the plausible values associated with the performance level for the booklet (ACT, 1997).

That difference could have accounted for the judgments indicating a relatively lower level of performance.

The current study design included the same panelists to classify the expected performance of their students *and* to classify the student booklets. Further, the special form of NAEP designed for the study included enough items to provide a reliable individual score estimate. Further design features in the selection of booklets helped to eliminate the effect of the NAEP policy of treating “not reached” items as “not administered.” Teacher panelists, in particular, have great difficulty in trying to take this policy into consideration when classifying test booklets according to achievement level categories.

As was the case for the 1995 studies, this study was conducted with grade 8 teachers and students. Results of grade 12 assessments with low stakes, such as NAEP, seem too problematic for a study of this importance. The curriculum for fourth graders is somewhat ambiguous in subjects such as U.S. history and civics. The Technical Advisory Committee on Standard Setting recommended that the study be conducted at grade 8. If resources permitted more than one grade, then grade 4 was the second choice. The study was implemented for grade 8 only.

This study not only informs the question of the validity of the Civics NAEP ALS process, but it also informs the debate regarding the use of item-by-item rating methods for the NAEP ALS process.

THE ASSESSMENT FORM USED FOR THIS STUDY

Selection of Item Blocks

A special form of the Civics NAEP was developed that would produce a reliable estimate of student ability in civics. The procedures used for selecting blocks of items for the 1995 study were replicated, as nearly as possible, for selecting items for the assessment in this study (Carlson, 1995). For the 1995 study, Educational Testing Service (ETS) selected the blocks according to technical criteria specified by ACT. The test form was to be constructed to include the minimum number of item blocks that could be selected to simultaneously satisfy the following criteria. This form should include items selected to maximize representation of the entire grade level item pool with respect to item difficulty, content (5 areas of civics instruction), and type of items (multiple choice and constructed response). Reliability should be approximately .90. ETS examined the information functions for the selected items, relative to the entire item pool. ETS determined that a “double-length” NAEP of approximately 100 minutes of testing time could be developed to meet the criteria. For the 1999 study, ACT assumed that four blocks would again be the minimum number that could meet this set of criteria for the study.

There are eight blocks of items for the 1998 grade 8 Civics NAEP, from which 70 distinct four-block combinations could be formed. Four blocks were selected to use in the study. Each set of four blocks was considered. Table 1 presents the criteria data for each of the 70 combinations. Darker shading identifies data that are within 1% of the grade-level item pool. The lighter shading identifies data that are within 2% of the grade-level statistics. The greater the number of darkly shaded cells, the closer the match of that set of blocks is to the grade-level item pool. Three combinations of four blocks were within 1% of the grade-level item pool for 10 of the 12 statistics in Table 1. The reliability of those three sets of four blocks each was computed (Hanson, 1999). The estimates of reliability for two of the three sets of four blocks were very close. The final decision regarding which set of item blocks to use was based on practical

considerations. The set of blocks requiring the fewest number of copyright permissions was chosen. Blocks 3, 5, 9, and 10 were used in the study. This set of four blocks had 75 items (several other combinations had 76). The percentage of extended constructed response items was slightly lower and the percentage of short constructed response items was slightly higher than that for the overall item pool. At the suggestion of the Technical Advisory Committee on Standard Setting (TACSS), information functions were plotted for the alternative sets of blocks. The items included in the SCS discriminate better at the upper scale scores and less well at the lower scale scores than the entire grade-level item pool. TACSS did not judge this difference to be significant.

Table 1
Selected Statistics of All Four-Block Combinations of 1998 NAEP Civics, Grade 8

Block	Number of Items									P-Value			
		Content Area					Type			Mean	Std	Min	Max
		1	2	3	4	5	MC	3	4				
All	151	19	35	44	22	31	123	22	6	49.62	17.99	12.60	89.40
3,4,5,6	76	12.58	23.18	29.14	14.57	20.53	81.46	14.57	3.97	48.42	17.15	12.60	85.60
		11.84	27.63	25.00	14.47	21.05	80.26	17.11	2.63				
3,4,5,7	76	9	19	20	12	16	61	12	3	46.57	18.93	12.60	85.60
		11.84	25.00	26.32	15.79	21.05	80.26	15.79	3.95				
3,4,5,8	76	8	20	20	12	16	62	11	3	48.42	17.71	12.60	86.80
		10.53	26.32	26.32	15.79	21.05	81.58	14.47	3.95				
3,4,5,9	76	9	19	20	12	16	62	11	3	47.59	18.16	12.60	89.40
		11.84	25.00	26.32	15.79	21.05	81.58	14.47	3.95				
3,4,5,10	75	8	20	21	11	15	61	11	3	46.66	18.58	12.60	85.60
		10.67	26.67	28.00	14.67	20.00	81.33	14.67	4.00				
3,4,6,7	76	9	21	20	11	15	61	12	3	48.70	18.24	12.60	85.60
		11.84	27.63	26.32	14.47	19.74	80.26	15.79	3.95				
3,4,6,8	76	8	22	20	11	15	62	11	3	50.55	17.03	12.60	86.80
		10.53	28.95	26.32	14.47	19.74	81.58	14.47	3.95				
3,4,6,9	76	9	21	20	11	15	62	11	3	49.73	17.47	12.60	89.40
		11.84	27.63	26.32	14.47	19.74	81.58	14.47	3.95				
3,4,6,10	75	8	22	21	10	14	61	11	3	48.79	17.89	12.60	85.60
		10.67	29.33	28.00	13.33	18.67	81.33	14.67	4.00				
3,4,7,8	76	8	20	21	12	15	62	10	4	48.70	18.81	12.60	86.80
		10.53	26.32	27.63	15.79	19.74	81.58	13.16	5.26				
3,4,7,9	76	9	19	21	12	15	62	10	4	47.88	19.25	12.60	89.40
		11.84	25.00	27.63	15.79	19.74	81.58	13.16	5.26				
3,4,7,10	75	8	20	22	11	14	61	10	4	46.94	19.67	12.60	85.60
		10.67	26.67	29.33	14.67	18.67	81.33	13.33	5.33				
3,4,8,9	76	8	20	21	12	15	63	9	4	49.72	18.04	12.60	89.40
		10.53	26.32	27.63	15.79	19.74	82.89	11.84	5.26				
3,4,8,10	75	7	21	22	11	14	62	9	4	48.79	18.46	12.60	86.80
		9.33	28.00	29.33	14.67	18.67	82.67	12.00	5.33				
3,4,9,10	75	8	20	22	11	14	62	9	4	47.97	18.90	12.60	89.40
		10.67	26.67	29.33	14.67	18.67	82.67	12.00	5.33				
3,5,6,7	76	10	18	20	11	17	60	15	1	49.52	17.81	12.60	84.20
		13.16	23.68	26.32	14.47	22.37	78.95	19.74	1.32				
3,5,6,8	76	9	19	20	11	17	61	14	1	51.36	16.60	12.60	86.80
		11.84	25.00	26.32	14.47	22.37	80.26	18.42	1.32				
3,5,6,9	76	10	18	20	11	17	61	14	1	50.54	17.04	12.60	89.40
		13.16	23.68	26.32	14.47	22.37	80.26	18.42	1.32				

Block	Number of Items									P-Value			
	Content Area					Type				Mean	Std	Min	Max
	1	2	3	4	5	MC	3	4					
3,5,6,10	75	9	19	21	10	16	60	14	1				
		12.00	25.33	28.00	13.33	21.33	80.00	18.67	1.33	49.61	17.46	12.60	84.50
3,5,7,8	76	9	17	21	12	17	61	13	2				
		11.84	22.37	27.63	15.79	22.37	80.26	17.11	2.63	49.51	18.38	12.60	86.80
3,5,7,9	76	10	16	21	12	17	61	13	2				
		13.16	21.05	27.63	15.79	22.37	80.26	17.11	2.63	48.69	18.82	12.60	89.40
3,5,7,10	75	9	17	22	11	16	60	13	2				
		12.00	22.67	29.33	14.67	21.33	80.00	17.33	2.67	47.76	19.24	12.60	84.50
3,5,8,9	76	9	17	21	12	17	62	12	2				
		11.84	22.37	27.63	15.79	22.37	81.58	15.79	2.63	50.54	17.61	12.60	89.40
3,5,8,10	75	8	18	22	11	16	61	12	2				
		10.67	24.00	29.33	14.67	21.33	81.33	16.00	2.67	49.60	18.03	12.60	86.80
3,5,9,10	75	9	17	22	11	16	61	12	2				
		12.00	22.67	29.33	14.67	21.33	81.33	16.00	2.67	48.78	18.47	12.60	89.40
3,6,7,8	76	9	19	21	11	16	61	13	2				
		11.84	25.00	27.63	14.47	21.05	80.26	17.11	2.63	51.65	17.69	12.60	86.80
3,6,7,9	76	10	18	21	11	16	61	13	2				
		13.16	23.68	27.63	14.47	21.05	80.26	17.11	2.63	50.82	18.14	12.60	89.40
3,6,7,10	75	9	19	22	10	15	60	13	2				
		12.00	25.33	29.33	13.33	20.00	80.00	17.33	2.67	49.89	18.56	12.60	84.50
3,6,8,9	76	9	19	21	11	16	62	12	2				
		11.84	25.00	27.63	14.47	21.05	81.58	15.79	2.63	52.67	16.92	12.60	89.40
3,6,8,10	75	8	20	22	10	15	61	12	2				
		10.67	26.67	29.33	13.33	20.00	81.33	16.00	2.67	51.74	17.34	12.60	86.80
3,6,9,10	75	9	19	22	10	15	61	12	2				
		12.00	25.33	29.33	13.33	20.00	81.33	16.00	2.67	50.91	17.79	12.60	89.40
3,7,8,9	76	9	17	22	12	16	62	11	3				
		11.84	22.37	28.95	15.79	21.05	81.58	14.47	3.95	50.82	18.70	12.60	89.40
3,7,8,10	75	8	18	23	11	15	61	11	3				
		10.67	24.00	30.67	14.67	20.00	81.33	14.67	4.00	49.89	19.12	12.60	86.80
3,7,9,10	75	9	17	23	11	15	61	11	3				
		12.00	22.67	30.67	14.67	20.00	81.33	14.67	4.00	49.06	19.57	12.60	89.40
3,8,9,10	75	8	18	23	11	15	62	10	3				
		10.67	24.00	30.67	14.67	20.00	82.67	13.33	4.00	50.91	18.35	12.60	89.40
4,5,6,7	76	11	17	21	11	16	61	12	3				
		14.47	22.37	27.63	14.47	21.05	80.26	15.79	3.95	48.29	17.25	16.00	85.60
4,5,6,8	76	10	18	21	11	16	62	11	3				
		13.16	23.68	27.63	14.47	21.05	81.58	14.47	3.95	50.14	16.03	16.00	86.80
4,5,6,9	76	11	17	21	11	16	62	11	3				
		14.47	22.37	27.63	14.47	21.05	81.58	14.47	3.95	49.32	16.48	16.00	89.40
4,5,6,10	75	10	18	22	10	15	61	11	3				
		13.33	24.00	29.33	13.33	20.00	81.33	14.67	4.00	48.38	16.90	16.00	85.60
4,5,7,8	76	10	16	22	12	16	62	10	4				
		13.16	21.05	28.95	15.79	21.05	81.58	13.16	5.26	48.29	17.81	16.00	86.80
4,5,7,9	76	11	15	22	12	16	62	10	4				
		14.47	19.74	28.95	15.79	21.05	81.58	13.16	5.26	47.47	18.26	16.00	89.40
4,5,7,10	75	10	16	23	11	15	61	10	4				
		13.33	21.33	30.67	14.67	20.00	81.33	13.33	5.33	46.53	18.68	16.00	85.60
4,5,8,9	76	10	16	22	12	16	63	9	4				
		13.16	21.05	28.95	15.79	21.05	82.89	11.84	5.26	49.31	17.04	16.00	89.40
4,5,8,10	75	9	17	23	11	15	62	9	4				

Block	Number of Items									P-Value			
	Content Area					Type				Mean	Std	Min	Max
	1	2	3	4	5	MC	3	4					
4,5,9,10	75	12.00	22.67	30.67	14.67	20.00	82.67	12.00	5.33	48.38	17.46	16.00	86.80
		10	16	23	11	15	62	9	4				
4,6,7,8	76	13.33	21.33	30.67	14.67	20.00	82.67	12.00	5.33	47.56	17.91	16.00	89.40
		10	18	22	11	15	62	10	4				
4,6,7,9	76	13.16	23.68	28.95	14.47	19.74	81.58	13.16	5.26	50.42	17.12	16.30	86.80
		11	17	22	11	15	62	10	4				
4,6,7,10	75	14.47	22.37	28.95	14.47	19.74	81.58	13.16	5.26	49.60	17.57	16.30	89.40
		10	18	23	10	14	61	10	4				
4,6,8,9	76	13.33	24.00	30.67	13.33	18.67	81.33	13.33	5.33	48.67	17.99	16.30	85.60
		10	18	22	11	15	63	9	4				
4,6,8,10	75	13.16	23.68	28.95	14.47	19.74	82.89	11.84	5.26	51.45	16.35	16.30	89.40
		9	19	23	10	14	62	9	4				
4,6,9,10	75	12.00	25.33	30.67	13.33	18.67	82.67	12.00	5.33	50.51	16.77	16.30	86.80
		10	18	23	10	14	62	9	4				
4,7,8,9	76	13.33	24.00	30.67	13.33	18.67	82.67	12.00	5.33	49.69	17.22	16.30	89.40
		10	16	23	12	15	63	8	5				
4,7,8,10	75	13.16	21.05	30.26	15.79	19.74	82.89	10.53	6.58	49.60	18.13	16.30	89.40
		9	17	24	11	14	62	8	5				
4,7,9,10	75	12.00	22.67	32.00	14.67	18.67	82.67	10.67	6.67	48.66	18.55	16.30	86.80
		10	16	24	11	14	62	8	5				
4,8,9,10	75	13.33	21.33	32.00	14.67	18.67	82.67	10.67	6.67	47.84	19.00	16.30	89.40
		9	17	24	11	14	63	7	5				
5,6,7,8	76	12.00	22.67	32.00	14.67	18.67	84.00	9.33	6.67	49.69	17.78	16.30	89.40
		11	15	22	11	17	61	13	2				
5,6,7,9	76	14.47	19.74	28.95	14.47	22.37	80.26	17.11	2.63	51.24	16.69	16.00	86.80
		12	14	22	11	17	61	13	2				
5,6,7,10	75	15.79	18.42	28.95	14.47	22.37	80.26	17.11	2.63	50.41	17.14	16.00	89.40
		11	15	23	10	16	60	13	2				
5,6,8,9	76	14.67	20.00	30.67	13.33	21.33	80.00	17.33	2.67	49.48	17.56	16.00	84.50
		11	15	22	11	17	62	12	2				
5,6,8,10	75	14.47	19.74	28.95	14.47	22.37	81.58	15.79	2.63	52.26	15.92	16.00	89.40
		10	16	23	10	16	61	12	2				
5,6,9,10	75	13.33	21.33	30.67	13.33	21.33	81.33	16.00	2.67	51.33	16.34	16.00	86.80
		11	15	23	10	16	61	12	2				
5,7,8,9	76	14.67	20.00	30.67	13.33	21.33	81.33	16.00	2.67	50.50	16.79	16.00	89.40
		11	13	23	12	17	62	11	3				
5,7,8,10	75	14.47	17.11	30.26	15.79	22.37	81.58	14.47	3.95	50.41	17.70	16.00	89.40
		10	14	24	11	16	61	11	3				
5,7,9,10	75	13.33	18.67	32.00	14.67	21.33	81.33	14.67	4.00	49.48	18.12	16.00	86.80
		11	13	24	11	16	61	11	3				
5,8,9,10	75	14.67	17.33	32.00	14.67	21.33	81.33	14.67	4.00	48.65	18.57	16.00	89.40
		10	14	24	11	16	62	10	3				
6,7,8,9	76	13.33	18.67	32.00	14.67	21.33	82.67	13.33	4.00	50.50	17.35	16.00	89.40
		11	15	23	11	16	62	11	3				
6,7,8,10	75	14.47	19.74	30.26	14.47	21.05	81.58	14.47	3.95	52.54	17.01	18.00	89.40
		10	16	24	10	15	61	11	3				
6,7,9,10	75	13.33	21.33	32.00	13.33	20.00	81.33	14.67	4.00	51.61	17.43	18.00	86.80
		11	15	24	10	15	61	11	3				
		14.67	20.00	32.00	13.33	20.00	81.33	14.67	4.00	50.78	17.88	18.00	89.40

Block	Number of Items									P-Value			
		Content Area					Type			Mean	Std	Min	Max
		1	2	3	4	5	MC	3	4				
6,8,9,10	75	10	16	24	10	15	62	10	3	52.63	16.66	21.20	89.40
		13.33	21.33	32.00	13.33	20.00	82.67	13.33	4.00				
7,8,9,10	75	10	14	25	11	15	62	9	4				

Construction of the Forms

Two different forms were developed, and they were distributed equally to students in each class. The two forms included the same items, but the order of the cognitive item blocks was reversed. Two booklets were printed for each form and one form (two booklets) was distributed to each student. Each form contained 7 sections with four cognitive blocks (two in each booklet) and student information. Form A included item blocks arranged in ascending numerical order and Form B included item blocks arranged in descending numerical order. This design allows for examination of a fatigue effect. The second booklet for each student included the three sections to collect information from students on background demographic data, courses taken and topics studied in civics and social studies, and study behavior and test motivation.

In addition to administration of the special form of the NAEP, ACT also collected information on the School Questionnaire and the Teacher Questionnaire for each school and teacher included in the study. This information was collected to allow exploration of possible sources of differences, should there be sizable differences between the performance of these students and that of the national sample. Comparisons of students, teachers, and schools in the study with those in the NAEP are provided later in this report.

Sections 1-4 of form A contained blocks C, D, I, and J in order, followed by the three sections of student background, curricular, and motivation questions. Sections 1-4 of form B contained blocks J, I, D, and C in order, followed by the three sections of student background information. Three sections of student survey questions were always placed at the end of the last booklet in the forms because they were placed at the end of the regular NAEP form.

PLANNING AND LOGISTICS

Panelists had been told during the pilot study and the ALS that a validation study involving grade 8 teachers was planned. In January 1999 ACT contacted panelists and schools to secure their agreement to participate. Both teachers and school principals were contacted regarding their participation.

An important change was made in the design of the study sequence. The assessment of students would not occur at exactly the same time as the teachers' classifications of student performances. For the 1995 study, TACSS had recommended that having the teachers classify students first would introduce less potential jeopardy to the study than having students assessed first. That is, TACSS perceived less impact from teachers' informing students of the classifications than of students' informing teachers of their performance. Nonetheless, students had to be assessed first in order to use student booklets in the booklet classification study. The benefits of this study design greatly outweighed the potential costs of this change. Students were assessed before the teachers were convened for the classification studies. Teachers were aware of the fact that they were not to look at the test booklets and that they were not to discuss performances with students.

The goal was to assure that every detail of this assessment and administration was as similar to a “standard” NAEP as possible. ACT provided schools with a choice of letters to be used to secure parental permission for students to participate in the study. Copies of the form letters prepared by ETS for NAEP distribution were obtained for the 1995 study, and those were again used in this study. ETS supplied item data, items, and scoring rubrics. ACT selected the items, secured copyright permission for materials to be included in the test booklets, and designed the cover for the new test booklets. NCES had requested that the cover be clearly distinguishable from the 1998 Civics NAEP. NCS printed the booklets; packaged spiraled sets of them for each administration session scheduled; and shipped them to Westat field staff. Westat administrators contacted the schools and established the exact schedule for testing, as well as the location for administration. They worked out all remaining details with the schools, and they contacted ACT regarding any “nonstandard” requests or circumstances. Westat staff returned booklets to NCS for scoring. NCS provided ACT with preliminary data about 10 days prior to providing the final quality controlled data files. These preliminary data were needed for ACT to begin procedures scoring booklets and for sampling booklets to include in the booklet classification study.

In order to administer the assessment, it was necessary to obtain clearance from the Office of Management and the Budget (OMB). (See Appendix A.) The first announcement of intent to collect these data was submitted to the *Federal Register* on November 18, 1998. The second notice was posted on March 2, 1999, and the request for clearance was submitted to OMB at the same time. OMB gave clearance to NAGB on May 11, 1999. The clearance package was prepared by ACT and submitted to NAGB. The actual request and permission was a transaction between the government agencies.

TEST ADMINISTRATION

Westat prepared to administer the assessment during the first two weeks of May to students in 13 schools. The schools included in the study were those represented by grade 8 teachers who served on the panel for either the pilot study or the ALS process for the 1998 Civics NAEP. Twenty-eight teachers were eligible to participate in the study: 11 teachers in the grade 8 pilot study panel and 17 in the ALS panel. Only 13 had been able both to get agreement from their schools to have their students participate in the study and to meet the planned schedule for the study for administration and for the validation study panel meeting. Some teachers and schools were willing to have their students participate in the assessment, but the teachers were not available for the classification study. The panelists were originally asked to participate in the study scheduled for March, but that schedule had to be changed. The beginning of May was a feasible time for Westat administrators to help with the study, and 13 schools and teachers agreed to participate in the study at that time.

The clearance from OMB was delayed as a result of several factors. This caused a last-minute change in schedules for administration. Westat contacted each school and scheduled administration a few days later. Three of the schools could not accommodate the change in schedule. Some of them already had other assessments scheduled, and some were too near the end of their school year. Two of the schools *included* in the assessment ended their school year in May: May 25, and May 26 were the closing dates for those schools. The remaining schools ended the school year between June 4 and June 22.

Westat prepared instructional materials for the administrators and shared those with ACT for review and approval. In a letter to teachers, ACT explained that the study design would be

jeopardized if they saw the NAEP to be administered to their students. Teachers were also asked to avoid discussions with students about the test difficulty, their performance, and the particular items on the assessment. Administrators were given similar instructions. One teacher reported that an administrator had shared a copy of the test with her. ACT did not follow-up with Westat on that report, however.

Westat administrators reported no unusual circumstances or problems during the administration.

PARTICIPATION IN THE STUDY

TEACHER PANELISTS AND THEIR SCHOOLS

Eighth grade civics teachers who had served on either the pilot study or ALS panel were invited to participate in the validation study. A total of 28 teachers had participated in the process. Only 13 teachers were able to participate and to get agreement from their schools to have their students participate in the study. Due to delays in receiving the OMB Clearance and to the fact that the school term was ending very soon for some schools, only 11 schools were included in testing.

Teachers whose students could not be included in the assessment because of the delays and changes in scheduled test dates were invited to participate in the panel study. A total of 11 teachers participated as panel members for the study. Panelists included two teachers whose students were not tested; one teacher whose students were tested did not participate. Two of the teachers had been pilot study panelists, and the remaining 9 had been ALS panelists. The largest number of students tested for any one teacher was 57 and the smallest was 31. Only two validation study members were males. (Please refer to the Appendix B for additional information.)

Nominees whose credentials are the most outstanding are selected to be ALS Panelists. When compared to the national sample of grade 8 teachers, the outstanding quality of the teachers in this special study is evident. A higher percentage of the teachers in the study have taught in the subject for more years, have earned more academic degrees, and have engaged in more professional development than teachers in the national sample. They are far more likely to engage in innovative instructional and assessment techniques. Data from the **Teacher's Questionnaire** are reported in Table 1 of Appendix C.

The resources of schools participating in the special study seem to be higher than those in the national sample. All of the schools in the study have at least 26 computers in the school, and nearly two-thirds of them have 100 or more computers in the school. They all have full-time library staff. These schools are not without their problems, however. The data in Table 2 of Appendix C show that the principals/chief administrators of the schools in this study were more likely than the national sample to identify nine behaviors or conditions as at least a minor problem. Behaviors such teacher absenteeism, student misbehavior, gang activities, and lack of parent involvement were all considered to be at least a minor problem. Lack of parental involvement stood out as a more frequent problem among the schools in this study, and the responses were explored in greater detail. The following data elaborate on this problem area.

- 30% of the schools in the national sample report that *more than one-quarter* of their parents participate in parent-teacher organizations compared to only *10% of the schools in this study*.

- 61% of the schools in the national sample report that *more than half* of their parents participate in open house or back-to-school nights compared to *44% of the schools in this study*.
- 9% of the schools in the national sample report that *more than half* of their parents participate in volunteer programs compared with *none of the schools in this study*.
- 8% of the school in the national sample report that *more than 10%* of their parents serve as assistants in classrooms compared to *none of the school in this study*.

STUDENTS AND THEIR SCHOOLS

The schools participating in the study are located in the northeast (3), southeast (5), and central (3) NAEP regions. None of the teacher panelists from the western region could participate. One teacher panelist was from Guam, and he was not invited to participate because transportation costs for administration and for his participation were judged to be too great.

A total of 499 student records were produced. Thirty-six records had no response data, so they were excluded. In addition, one student did not return for the second half of the assessment and another was reported to have been absent for a significant portion of the test period. Both of those were excluded. Therefore, the total valid number of students assessed is 461.³

As reported above, one teacher did not participate in the classification study although her students were assessed. There were 30 students in that class, and those students were excluded from this report in order to maintain consistency in the number of student classifications.

In addition, the performance of 17 students was extremely low. The theta estimates for their performance was very low—lower than -6 . The Technical Advisory Committee on Standard Setting recommended elimination of student assessment records for which the theta estimate was lower than -3.0 because score estimates in that performance region are very unstable. The final count of students included in this study report is 414.⁴

The racial/ethnic identity of these students was somewhat different from that of the national sample. In particular, there were fewer white and African American students and more “Hispanic” students. Data for the national sample and the students in the study are reported in Table 2.

³ For details about the number of records, number of items, coding of missing data (omits and not reached), and an assortment of other topics, please see “Notes for Data Cleaning, File Management, and Scoring for the 1998 NAEP Civics Validation Study” by Yang in Appendix D.

⁴ Performances of the low performing students were analyzed extensively. Please refer to the Hanick paper in Appendix D for a report on performances of low performing students which includes data on teacher classifications and comments about their classification decisions.

Table 2
Racial/Ethnic Identity of Students in the National NAEP Sample
Compared to the Special Study Sample

Racial/Ethnic ID	National Sample %	Study Sample % (n=414)
White	67.1%	57.9%
Black	14.5	10.2
Hispanic	13.7	21.7
Asian/Pacific Islanders	3.4	5.6
American Indian/Alaska Natives	1.2	0.7

Coupled with the relatively higher racial/ethnic diversity for the students in this study (relative to the national sample for grade 8 in the Civics NAEP), a lower percentage of the students in this study are life-long U.S. residents (85% of the students in the study vs. 91% for the national sample) and a larger percentage are rather recent arrivals to this country (2.7% vs. 1% have lived in the US less than 3 years). A slightly higher percentage of the students in the study reported that a language *other than* English is spoken in the home all or most of the time (17% vs. 12%), and nearly a third of them reported that a language other than English is spoken in the home at least half of the time.

The educational attainment of the parents of these students is not very different from that of the national sample. About 43% of the students reported that their mothers had graduated from college, and 36% of them reported that their fathers had completed college. These figures for the national sample of students are 38% and 36%, respectively.

In terms of the civics topics studied, more students in these classes reported having studied the different topics than was the case for students in the national sample. Over 90% of these students reported having studied the US constitution and Congress, compared to only 75-80% of the national sample. Over 80% of these students reported studying the legislative process, the court system, political parties and the electoral process, and state and local governments, whereas only two-thirds or fewer of the students in the national sample studied these topics. About 80% of these students reported having studied the President and Executive branch, compared to about 55% of the national sample. Finally, less than half of these students reported having studied governments of other countries and international relations/organizations, compared to about one-third of the national sample.

Descriptive statistics for the performance of students in each school are reported below in Table 3. The theta values computed as the mean and median performance levels for the schools are reported, along with the numbers of students included in the computations and the minimum and maximum score values. The standard deviation reported is for the mean. The achievement level category associated with the mean and median scores is also noted. Both the mean and median performances for most schools were in the Basic range of achievement. The mean and median scores for students tested in one school were in the Proficient range, and the remaining eight were in the Basic range.

Table 3
Performance on the Special Form of the Civics NAEP
by Students Tested Whose Teacher Participated in the Classification Study

Mean Scores on the ACT NAEP-Like Scale for Students in this Report (n=414)	Median Scores on the ACT NAEP-Like Scale for Students in this Report (n=414)	N Students in Report from the School	Minimum Score	Maximum Score
161.02 (B)*	161.72 (B)	45	136.80	176.00
161.44 (B)	163.26 (B)	40	134.56	176.98
164.38 (B)	164.52 (B)	37	145.62	175.58
151.92 (B)	152.06 (B)	51	118.04	169.00
162.63 (B)	165.50 (B)	43	127.42	182.86
174.18 (P)	173.06 (P)	62	158.50	184.68
150.24 (B)	152.06 (B)	35	116.36	169.00
155.00 (B)	158.78 (B)	59	121.82	179.92
157.03 (B)	156.96 (B)	42	129.10	173.06

*B = score was within range of Basic cutscores. P = score was within range of Proficient cutscores. No school had a mean or median score at or above the Advanced cutscore.

Only three schools had students scoring in the Advanced score range, while eight schools had students with scores below the Basic level. Only one school, the school with the mean and median performance levels in the Proficient range, had no student scoring below the Basic achievement level; all others had at least one student scoring at the Below Basic level.

Please note that one school included in the special assessment included mostly seventh grade students. The fact that the teacher had almost no eighth-grade civics students was revealed at no time during the ALS process nor during the validation study. This fact was revealed when ACT staff read the comments of this teacher regarding the reasons for which students were classified at specific levels of achievement. The teacher's comments revealed that 35 of the 37 students included in the assessment were 7th grade students with little civics instruction. He also indicated that seven of those students were repeating 7th grade—one for the third time. In addition, two 8th grade students in the study were reported to be repeating the social studies course for the third time.

The mean performance level (150) for the 7th grade students was the lowest of any of the 9 schools reported. The median performance level was tied at 152 with one other school. The performance of the students from this school was analyzed relative to that of all other students in the study. In general, the students appeared to be low performing students. Written comments by the teacher indicated that 15 students had very low grades in all classes, and two of the students were getting straight F's. The teacher also reported that five of the students were "borderline Special Ed" students. One of the five actually qualified for Special Education Classes, but his parents refused the placement.

Aside from low performance, the major difference between these students and other students in the study seems to be that these students did not study topics in civics. It is a bit difficult to understand how some students in a course reported that they had studied a topic and some reported that they had not. Whatever the explanation, a smaller proportion of these students reported having studied the various topics in civics and larger proportions responded to the

questions with “Don’t Know.” These responses indicate that the students were, perhaps, such low ability students that they actually did not know what they had studied.

Teachers are not required to teach a civics course, per se, but the guidelines for participation on a panel require that the teacher panelists teach a social studies course for which civics content is a significant component. That guideline does not appear to have been met in this case. This teacher commented that a lack of civics would hold four [of his/her better] students back on the test.

COMPUTATION OF STUDENT SCORES

ACT planned to compute student scores using both a maximum likelihood estimation procedure and a Bayesian EAP procedure. The maximum likelihood estimates would be used for analyses and reporting, and the EAP scores would be computed as a quality control check for the computations.⁵ ACT attempted to replicate the computational procedures used by ETS as nearly as possible. The software that is available commercially for computing estimates of student performance scores is not the same as that used by ETS for computing NAEP scores, however. When PARSCALE, the commercially available software, was used to estimate the theta values for student performance, no estimates were obtained for an extraordinarily large number of cases. ACT was unable to resolve this problem with the software and unable to determine whether missing data were, in fact, being handled in the desired manner with the analysis program. ACT was able to use PARSCALE to produce EAP estimates for all the student records. Ultimately, a program was written to produce EAP and likelihood estimates of student scores. Those scores were checked against the output of PARSCALE and found to be almost exactly the same. Extensive analyses were conducted on the data. The Technical Advisor Team evaluated the results, and the computational procedures were judged to be sound and accurate.

For purposes of selecting student booklets for the Booklet Classification portion of the study, EAP score estimates were used. The computational program for the EAP procedure had already been prepared to serve as the quality control check on maximum likelihood estimates (MLE). Because there was limited time for selecting booklets and preparing them for the study, these EAP estimates were used in the booklet selection process. In order to produce results as similar as possible to those expected from a maximum likelihood estimation procedure, the values of the parameters set for the computational procedure were altered to represent a very weak set of assumptions. The computations were first made using a Normal prior distribution with mean = 0 and standard deviation = 1. The value of the standard deviation was changed to 5 to assure that the distribution used as the prior for computations was relatively flat over the range of interest and was reflective of a “non-informative” prior. (Please refer to the July 6, 1999 Scoring Memo in Appendix D for more details about these procedures.) Members of the Technical Advisory Team again evaluated the procedures and results, and they judged them to be sound and accurate.

Once the booklets were selected, ACT staff developed a maximum-likelihood computational procedure for estimating scores. Those results were compared to the EAP estimates for all student booklets and for the booklets used in the Booklet Classification portion of the study. The Pearson correlation between the two sets of estimates was .99. A plot of the two sets of estimates is included as Figure 1 in Yang (1999) in Appendix D. While the achievement level classification of eight student performances differed between the two computational procedures, none of the 50

⁵ A report by Scoring included in Appendix D details the estimation procedures that were used by ACT.

booklets in the Booklet Classification Study (10 for practice and 40 for the classifications) changed. To a great extent, this was a result of the recommendation of TACSS to select booklets that were not near borderlines of the achievement levels. All eight of the different classifications in the study resulted from higher likelihood estimates relative to the EAP estimates. Performances of five students scored in the Proficient range by the MLE procedure were scored in the Basic range by the EAP estimates. Three performances scored within the Basic range by the MLE procedure were scored in the below Basic range by the EAP estimates. The likelihood estimates were used for this study to determine the “empirical score” classifications.

THE STUDY

The eleven teachers were convened in St. Louis at the Ritz Carlton Hotel July 9-11, 1999. A copy of the agenda is included in Appendix E.

Panelists arrived on Thursday afternoon and checked in with ACT staff. The meeting began at 8:30 AM on Friday, July 9. Panelists were seated at round tables with 3 persons per table for three tables and 2 persons at the fourth table. Seating charts from the previous meetings were reviewed, and all validation study panelists were assigned to table groups that included no former tablemates.

An overview of the validation study was presented to panelists: the design, the purpose, and the importance of conducting validation studies. Panelists were given an opportunity to ask questions about the study.

Next, an update and overview of the process was presented to help panelists recall the procedures they had used for developing achievement levels. Data from Round 3 and the final cutpoints resulting from the achievement levels-setting process were shared with panelists. A brief description of subsequent studies and analyses that had been conducted was presented to the panelists. The study facilitator made it clear that ACT had analyzed many, many aspects of the ALS process and that the results of those analyses had been shared with the Technical Advisory Committee on Standard Setting. The facilitator also made it clear to the panelists that the National Assessment Governing Board had accepted the recommendations from ACT and that ACT had recommended the results they had developed in the ALS process.

Panelists were re-trained in the Framework and the ALDs by the content expert who had worked with them previously. Exercises to help panelists become “recalibrated” with respect to performance levels and achievement levels descriptions were also included as training for each step in the process. Panelists were given a set of exemplar items and student papers to review as part of their recalibration process.

Teachers had been told to bring their grade books to the meeting or to review them carefully before coming. This change in procedures was introduced as a result of experiences in the 1995 SCS. Teachers in the geography and U.S. history study indicated that they would have benefited from having their grade books available to refresh their memories of student performances in their classes. They were instructed in the use of their grade books and cautioned not to confuse course grades with performance relative to the achievement levels descriptions.

Panelists were asked to complete four process questionnaires administered at different times throughout the process.

Implementation of the Similarities Classification Study

Classification of the Overall Level of Civics Knowledge and Skills

The first part of the study is called the Similarities Classification Study (SCS), and it included two different classifications. Teachers were asked to give classifications for each of their students included in the classes that had participated in the special NAEP assessment. The name of each student in the assessment was printed on a classification form for each teacher. A copy of a classification form is included in Appendix E.

Teachers were first asked to classify the overall achievement of civics knowledge and skills for each student, according to the grade 8 achievement levels descriptions. Panelists were instructed to mark the location on the form that best corresponds to each student's level of knowledge and skills in civics, relative to the achievement levels descriptions. They were instructed to base their classifications on the achievement levels descriptions. Teachers were also asked to rate their confidence in their judgment of the achievement level classification of each student. Confidence ratings were simply *low*, *medium*, or *high*.

They were to mark their classification of each student on the scale printed on the form for each student. A total of ten locations were identified for marking: solid Below Basic, upper Below Basic, lower Basic, solid Basic, upper Basic, and so forth through solid Advanced (no upper Advanced). In addition to marking the location to represent student performance relative to the achievement levels, panelists were also asked to mark their level of confidence (high, medium, or low) regarding the achievement level classification for each student. Please see appended material in Appendix E.

Teachers performed this task much faster than anticipated. All had finished within about one hour. Panelists completed the first process evaluation questionnaire at the end of that session.

An On-Site Change in the Study Design: Round 2 of Classifying the Overall Level of Civics Knowledge and Skills

During the first classification of their students, teachers commented on how their students would perform on NAEP items and some other aspects of student test-taking behavior. The facilitator cautioned them each time to focus only on the ALDs and their judgment of each student's knowledge and skills in civics.

Observers from the NAGB staff recommended a change in the study design to collect more information. They suggested that a sort of replication study be conducted before the second planned classification of students. This replication was to determine whether the panelists changed their classifications of students and to collect information for each classification regarding factors that potentially influenced teacher's classifications.

NAGB staff worked on the list of factors and the format of the questions, but they ultimately left this task to the study facilitator and coordinator. Student files were transmitted electronically, and new classification forms were developed for the second classification of student achievement with respect to their overall civics knowledge and skills.

Half of the students (every other student on the roster) for each teacher were included in the study. A copy of the classification form is included in Appendix E. Teachers were asked to rate each of six factors that "could have influenced your classification of this student." They were

given a Likert-type scale (5 = *very large influence*; 3 = *some influence*; 1 = *no influence*) for responding with regard to the following factors:

1. Overall knowledge and skills in all subjects
2. Overall knowledge and skills in civics
3. Test-taking behavior
4. Achievement levels descriptions
5. Items on the Civics NAEP
6. Grade(s) in my course

The panelists were told that this extra classification was added for purposes of collecting more information regarding the classification process. This addition was, in part, made because the amount of time required for the classifications was considerably less than had been scheduled. The panelists did not appear to be bothered by this additional task. This second round of the first classification procedure required almost twice as much time as the first. A copy of this classification form is included in Appendix E.

During the process of classifying half of their students in the second round of “overall civics knowledge and skills” judgments, teachers commented that having the list of factors was causing them to think about those factors now, although they had not taken them into account during the first round. The results of this classification, therefore, seem contaminated, and ACT recommends that little attention be given to these classifications.

Analyses were conducted to determine whether these factors appeared to have influenced comments collected in the second, planned classification in the study. There was no evidence that this was the case. Perhaps the difference in the two classification tasks in the SCS accounts for the fact that there was no apparent influence from these factors on the classifications collected for students’ expected performance on the special form of NAEP. That is, the factors influencing the teachers’ classifications of the overall knowledge and skills of their students appear to be quite different from the more individualized factors influencing classifications of expected student performance on the special form of NAEP.

Classification of Expected Student Performance on the Special Civics NAEP

Teachers were asked to classify each student’s expected performance on the special Civics NAEP that had been administered. Before beginning the classification process, however, teachers were engaged in training exercises. These exercises included review and discussion of ten student booklets that were pre-classified as being within one of the three achievement levels categories or at the Below Basic level. They were also given item booklets to review. These item booklets included all items in the grade 8 Civics NAEP and the scoring rubrics for each. In addition, they were given their own Reckase Charts⁶ from Round 2 of the ALS process in which they had participated. They could look at their ratings for specific items. The purpose of these exercises was to help panelists be calibrated with respect to the cutscores set for each achievement level. These reviews helped panelists focus on the level of item difficulty in the assessment and how the difficulty, along with the content of the item, related to the achievement levels descriptions.

Teachers were instructed in the criteria used to select items for the special form of NAEP. They were told that the items met these criteria and that the assessment of cognitive items lasted 100 minutes. They were also told that students had a break after 50 minutes of testing. They were not

⁶ Please refer to Loomis (2000) and Loomis & Hanick (2000) for more information about Reckase Charts.

told which items were included on the special assessment. They were instructed to again rate their confidence in their judgment of the achievement level classification of each student. Finally, they were instructed to comment on the factor(s) taken into account in classifying the expected performance of each student. (Please refer to the instructions included in Appendix E.)

Teachers seemed to understand instructions and procedures well. They completed two process evaluations during the first part of the study. The first was completed at the end of Day 1, and the second was completed at the end of the classification of their students' expected performance on the special form of the Civics NAEP. Panelists' responses to question about the process are analyzed later in this report.

Findings of the Similarities Classification Study

Overall Knowledge and Skill in Civics

Data reported in the tables of results are averages of the percentages of students classified in each achievement level category by each individual teacher. The data tables were prepared in the same way as those for the 1995 study. A "hit rate" was again computed for each table. This value is the percentage of classifications by teachers that correspond to the student score classifications into achievement level categories (empirical score classification). P_A reports the hit rate, and it is the sum of the percentages of students for which the teacher classifications correspond to the empirical score classification.

Results reported in Table 4 indicate that the teachers were somewhat more likely to classify their students' level of overall civics achievement at a higher level than the student's empirical performance. These findings are the same as for geography and US history. That is, teachers tend to think that their students' achievement is as high or higher than the actual performance of their students on the special form of NAEP. The comparison in Table 4 is between the level at which teachers classified their students' overall knowledge and skills in civics and the level at which the student scores were classified, i.e., the empirical classification.

Teachers' classifications of the overall civics knowledge and skills of their students agreed with the students' empirical performance level in 44% of the cases. The association between score estimates of the achievement level performance categories of students and teachers' classifications of the overall knowledge and skills of the students is positive, but somewhat low ($K = .24$). Teachers tend to estimate the civics knowledge and skill achievement level category of their students to be higher than that derived for students on the basis of their performance on the special form of NAEP. This is evident in the relatively larger percentages in the cells above the diagonal of values in Table 4 compared to the percentages in the cells below the diagonal in Table 4. Teachers' classifications of overall civics knowledge and skills were within one achievement level (+/-) of the empirical level for 95% of the students, and they were the same or one level higher in 87% of the cases. Teachers classified the overall civics knowledge and skills of their students at a higher level than the empirical achievement level for an average of 47% of the students and at a lower level than the empirical achievement level for an average of only 13% of the students. This finding is similar to that for geography and US history in 1995.

Table 4
Percentage of Students Classified within Achievement Level Categories Based on Overall Civics Knowledge and Skills Relative to the Empirical (MLE) Score Classifications

Table N = 414	Achievement Level Classification of Overall Civics Knowledge and Skills (SCS#1)			
Achievement Level Classification of MLE Score Estimates of Student Performance (ACT NAEP-Like Cutscores)	Below Basic (n=72)	Basic (n=118)	Proficient (n=119)	Advanced (n=105)
Below Basic (<149.2) (n=64)	8.7% (n=36)	5.6% (n=23)	1.2% (n=5)	0.0 (n=0)
Basic (149.2 – 165.39) (n=189)	8.2 (n=34)	19.1 (n=79)	15.0 (n=62)	3.4 (n=14)
Proficient (165.4 – 177.89) (n=140)	0.5 (n=2)	3.9 (n=16)	11.8 (n=49)	17.6 (n=73)
Advanced (≥ 177.9) (n=21)	0.0 (n=0)	0.0 (n=0)	0.7 (n=3)	4.4 (n=18)

Bold entries are for cells that would represent “hits” or agreement.

P_A = .440
P_E = .267
K = .243

Of the 64 students who scored in the Below Basic category, teachers classified 36 of them in that category. This means that 56% of the students scoring below the Basic cutscore were also classified at that level by their teachers who rated their overall knowledge and skills in civics at the Below Basic level. Of course only half of the 72 students classified at the Below Basic level by their teachers actually scored at that level.

Of 21 students who actually scored at or above the Advanced level, teachers classified 18 of them as Advanced with respect to their overall knowledge and skill level in civics. There is no higher level at which students could be classified. That upper limit, coupled with the tendency for teachers to classify their students at a higher level, makes this higher correspondence (86%) rather unexceptional. Even more important is the fact that teachers classified 105 students in the Advanced category with respect to their overall knowledge and skills in civics. Seventy percent of those students actually scored in the Proficient score range, and 13% scored in the Basic level.

Teachers’ judgments of students corresponded with the student performance scores for about 42% of the students (79 of 189) scoring in the Basic achievement level range. Of the students (118) that teachers judged to be within the Basic level with respect to their overall knowledge and skills in civics, about two-thirds (79) actually scored within that range of performance. This compares to only about 41% of the students that teachers judged to be within the Proficient level (with respect to overall knowledge and skills in civics) that actually scored at that level (49 of 119).

Achievement levels descriptions were used to set the cutscores that served to classify student scores on the special form of the NAEP. Teachers who participated in that process used these descriptions to classify the overall knowledge and skills of their students in civics. Perhaps the relatively low correspondence between the two classifications is a result of the fact that teachers were asked to judge “overall” knowledge and skills in civics. Factors other than those included in the achievement levels descriptions are likely to be taken into account in making this judgment

about specific students that the teachers know personally. The achievement levels descriptions served to filter these factors when teachers were asked to characterize their students' overall knowledge and skills with respect to the achievement levels, but the two factors being compared here (overall knowledge and skills versus actual performance on a special form of the Civics NAEP) are rather distant. That distance was greatly decreased in the second set of judgments teachers were asked to make in the Similarities Classification Study.

Classification of Students' Expected Performance on the Special Form of NAEP

For the second set of classifications, teachers were asked to classify each student according to the achievement level category that would be expected for the student's performance on the special form of the Civics NAEP. Teachers were familiar with all the items in the grade 8 NAEP item pool, but they did not know which specific items were included on the special form used to assess their students. They knew the length of the assessment and the circumstances of the administration.

A pattern of findings similar to that reported in Table 5 is found in Table 5 for the relationship between teachers' expectations of the performance of their students on the special form of NAEP and the actual performance of their students on the assessment. The correspondence between actual performance on the special form and teachers' estimates of student performance was expected to be somewhat higher than that for estimates of overall knowledge and skills. Although teachers did not know the exact contents of the assessment, it seemed likely that the classification of expected performance would be more similar to actual student performance than the classification of overall knowledge and skills. Teachers' classifications of the expected performance of their students on the special form of NAEP agreed with the empirical performance level in about 44% of the cases. The value of the Kappa statistic is approximately the same as that for Table 4, and this again shows a low association between the two classifications.

Teacher's classifications of the expected performance of their students were within one achievement level of the students' empirical performance level in 97% of the cases. Overall, teachers classified the expected performance of 39% of their students at a higher achievement level than the empirical performance level, and they classified the expected performance on only 17% of their students as lower than the empirical level. Teachers classified expected student performance at a higher level than the empirical classification almost as frequently as they classified the performance in the same category as the empirical score classification.

Table 5
Percentage of Students Classified within Achievement Level Categories Based on
Expected Performance on the Special Form of the Civics NAEP
Relative to the Empirical (MLE) Score Classifications

Table N = 414	Achievement Level Classification of Expected Student Performance on the Special Form of the Civics NAEP (SCS#2)			
Achievement Level Classification of MLE Score Estimates of Student Performance (ACT NAEP-Like Cutscores)	Below Basic (n=87)	Basic (n=115)	Proficient (n=119)	Advanced (n=93)
Below Basic (<149.2) (n=64)	9.7% (n=40)	5.3% (n=22)	0.5% (n=2)	0.0 (n=0)
Basic (149.2 – 165.39) (n=189)	10.9 (n=45)	17.9 (n=74)	15.0 (n=62)	1.9 (n=8)
Proficient (165.4 – 177.89) (n=140)	0.5 (n=2)	4.6 (n=19)	12.6 (n=52)	16.2 (n=67)
Advanced (≥ 177.9) (n=21)	0.0 (n=0)	0.0 (n=0)	0.7 (n=3)	4.4 (n=18)

Bold entries are for cells that would represent “hits” or agreement.

P_A = .446
P_E = .268
K = .243

Table 6 reports the correspondence between teacher’s classifications of students with respect to the student’s overall knowledge and skills in civics and the student’s expected performance on the special form of NAEP. Overall, teachers tended to classify the overall civics knowledge and skills of their students as highest, and their expected performance as next highest. The two classifications corresponded for 73% of the students, and the classification of overall civics knowledge and skills was higher than expected performance on the special form of NAEP for 18% of the students. For only 9% of the students did teachers classify the expected performance on the special NAEP as higher than the overall knowledge and skills in civics.

Teachers’ classifications of their students with respect to overall civics knowledge and skills were more similar to their classifications with respect to expected performance on the special form of NAEP than either of these classifications was to the empirical performance classifications.

Table 7 provides summary data for the results of the classifications by teachers of the overall civics knowledge and skills of their students (SCS#1), by teachers of the expected performance of their students on the special form of NAEP (SCS#2), and summary data of the cutscore classifications based on the MLE score estimates (empirical performance scores levels). The percentages of students classified at each achievement level category by teachers are very similar for the two classifications teachers made. There were more students classified at the below Basic level and fewer at the Advanced level when expected performance on the special form of the Civics NAEP was the frame of reference than when overall knowledge and skills was used for the classification. Relative to the empirical score classifications, the proportions classified by both sets of teacher judgments in the Below Basic level and the Advanced level look quite high. Relative to the empirical score classifications, the proportions classified in the Basic and Proficient levels look quite low.

Table 6
Percentage of Students Classified within Achievement Level Categories Based on Overall Civics Knowledge and Skills (SCS#1) by Expected Performance on the Special NAEP (SCS#2)

Table N = 414	Achievement Level Classification of Overall Civics Knowledge and Skills (SCS#1)			
Achievement Level Classification of Expected Performance on Special Form of Civics NAEP (SCS#2)	Below Basic (n=72)	Basic (n=118)	Proficient (n=119)	Advanced (n=105)
Below Basic (n=87)	14.0% (n=58)	6.5% (n=27)	0.5% (n=2)	(n=0)
Basic (n=115)	3.4 (n=14)	18.6 (n=77)	5.6 (n=23)	0.2 (n=1)
Proficient (n=119)	(n=0)	3.4 (n=14)	20.1 (n=83)	5.3 (n=22)
Advanced (n=93)	(n=0)	(n=0)	2.7 (n=11)	19.8 (n=82)

Bold entries are for cells that would represent “hits” or agreement.

Table 7
Percentage of Students Classified within Achievement Level Categories Based on Overall Civics Knowledge and Skills (SCS#1) and Expected Performance on the Special NAEP (SCS#2) Relative to Percentages Based on MLE Scores Within Cutscore Ranges for each Achievement Level (Empirical Score Level)

	SCS #1 % Classified Within Level (n=414)	SCS #2 % Classified Within Level (n=414)	% Within Empirical Score Level (N=414)
Below Basic	17.4% (72)*	21.0% (87)	15.5% (64)
Basic	28.5 (118)	27.8 (115)	45.7 (189)
Proficient	28.7 (119)	28.7 (119)	33.8 (140)
Advanced	25.4 (105)	22.5 (93)	5.1 (21)

*Numbers in parentheses are the numbers of students classified in each achievement level category.

Table 8 reports the mean and median values of MLE empirical score estimates, reported on the ACT NAEP-like score metric, of students who were classified according to each of the classification criteria. These data help to provide more understanding of the level of similarity in the performance of students classified for each classification condition. For example, the scores on the special form of NAEP were considerably higher for students classified by their teachers at the Below Basic level than the scores that were classified according to the achievement level cutscores. Teachers generally classified their students at a higher level than the empirical score classification level. The data reported in Table 8 seem to be generally consistent with that finding. The scores of students in the classification levels determined by teacher judgments were

generally lower than those classified by empirical score levels. At the Below Basic level, however, the mean and median scores of students classified by their teachers are higher than those classified according to the empirical score level.

Table 8
Mean and Median ACT NAEP-Like Score Values for MLE Score Estimates of Performance by Students: Three Classifications of Performance

Achievement Level Classification of MLE Score Estimates of Student Performance (ACT NAEP-Like Cutscores) (Table n = 414)		Overall Knowledge and Skills	Expected Performance on Special Form of Civics NAEP	Empirical Score by Achievement Level
Below Basic (<149.2)	Mean	146.18	147.30	137.92
	Median	148.84	149.82	140.58
	n =	(72)	(87)	(64)
Basic (149.2 – 165.39)	Mean	154.72	155.42	157.52
	Median	156.40	156.82	157.8
	n =	(118)	(115)	(189)
Proficient (165.4 – 177.89)	Mean	163.54	163.96	170.68
	Median	164.24	164.94	170.4
	n =	(119)	(119)	(140)
Advanced (≥177.9)	Mean	172.08	173.20	181.18
	Median	172.08	172.92	181.18
	n =	(105)	(93)	(21)

*Numbers in parentheses are the numbers of students classified in each category.

SUMMARY OF FINDINGS

Findings from the 1999 study for civics were very similar to those for the 1995 study for geography and US history. Results indicate that the teachers were somewhat more likely to classify their students' level of overall civics achievement and expected performance on the special NAEP at a higher level than the student's empirical performance. *Based solely on these findings, one would conclude that achievement levels for NAEP were set too high.*

BOOKLET CLASSIFICATION STUDY

The 1999 validation study for civics combined the two separate studies that had been implemented in 1995 for geography and U.S. history (the SCS and the BCS) into a single, comprehensive study. The rationale of the study was described earlier, but a review might be

worthwhile before beginning to analyze the booklet classification study. The combination of the designs of two studies that had been implemented in 1995 provided the opportunity to address some questions that were raised then. The results of the two studies pointed to contrary conclusions. When teachers classified the overall knowledge and skills of their students and the expected performance of their students, they tended to classify them at a higher level than the actual performance of the students. When panelists classified booklets representing student performances on NAEP, they tended to classify them at a lower level than the actual performance of the students. The first result would suggest that the achievement levels were set too high, and the second would suggest that the levels were not set high enough. It was not clear whether these were general findings or findings due to differences in panel members for the two studies or differences related to the use of regular NAEP booklets, an so forth.

The fact that the same panelists were included in both the SCS and BCS for civics was judged to be a positive and significant change in the study design. Panelists for previous booklet classification studies had included general public representatives and educators who were not K-12 classroom teachers, as well as teachers who met the same criteria as the teacher panelists included in this and other ALS studies. Only teachers participated in the BCS for civics. No consistent pattern had been discerned in previous studies to indicate that teachers would classify the booklets higher or lower than other types of panelists.

SELECTION OF BOOKLETS FOR THE CLASSIFICATION STUDY

A total of 50 booklets were to be selected from the entire set of students who participated in the study. Forty booklets scored within the range of each achievement level category were selected for the classification study and 10 for the practice session. The guidelines recommended by TACSS for selecting booklets were as follows:

- Distribute booklets in each achievement category to include 7 in the below Basic range, 13 in Basic, 13 in Proficient, and 7 in Advanced.
- Select booklets with scores that are 1 or 2 standard errors from the cutpoints. The decision was made to use the Brennan standard error of the cutscores, which is generally quite low. Booklets with scores at least 2 standard errors from the cutscores were included in the set from which study booklets were selected.
- Include at least one student from each school in the study.
- Include no more than 5 – 7 students from any one school.
- Include students from at least two different schools in each achievement level category.
- Do not include a booklet if a “significant number” of items were left blank. In particular, do not include booklets for which missing data would be considered “not reached,” i.e., not administered.

During the process for selecting booklets, some additional guidelines were developed and implemented. Only the booklets with scores that were at least two standard errors from the cutscores were included on the list for selection, and they were ordered from lowest to highest scores. The number of booklets in each achievement level category was counted and divided by the number to be selected. The quotient (n) was used to identify the booklets to select: every n^{th} booklet was selected in each achievement level range. Neither the lowest nor the highest (first nor last) booklet score in the range was selected. In order to include booklets in each achievement category from as many different schools as possible, some booklets were substituted for the n^{th} booklets. The substitutions were as close to the n^{th} booklet as possible. Further,

booklets within the same school were substituted in order to equalize the number of booklets from the two forms (A and B) included in the study.

The booklet selections were made on the basis of the EAP score estimates. Once the maximum likelihood score estimates were computed, comparisons were made of the achievement level classification of each booklet, based on the two computational procedures. The empirical score classifications for the two procedures were the same for all booklets used in the practice and in the classification for the study. A copy of the comparison table is included in Appendix D as Table 3.

TRAINING PANELISTS FOR THE BOOKLET CLASSIFICATIONS

For the second part of the study, teachers were asked to use the ALDs as the criterion for classifying performances represented by student test booklets. Training for this procedure began after lunch on the second day. Teachers were told that there would be 40 booklets in the Booklet Classification Study (BCS). They were told only that booklets were selected such that the score of at least one booklet fell within the range of each achievement level. The scores of the booklets were not revealed to the teachers, nor were the empirical achievement level classifications. Individual item scores were not revealed to panelists. Panelists had scoring rubrics for all items, and they could refer to those rubrics.

Panelists were given a form to use for marking the achievement level category in which each booklet was classified. Four achievement level categories were available for the panelists to select. Booklets were numbered from 1-40, and the numbering was unrelated to the score of the booklet and unrelated to the school identification.

THE PRACTICE SESSION OF BOOKLET CLASSIFICATIONS

Panelists were instructed in the method and in marking their classification forms. In particular, they were told that they were to classify the booklets according to the achievement levels descriptions and to base their classifications on a holistic judgment. The facilitator stressed that scoring booklets was not the task and that booklet scores were not necessary in order to perform the task.

Panelists were given 10 booklets to classify in a practice session. They were given one hour for this practice classification. They were told that the rate (10 booklets per hour) would be the approximate rate necessary for the actual BCS the following day. All panelists completed classification of all booklets in the practice set within the allotted time. Several just managed to complete the task in that time, and most seemed somewhat concerned with the pace that would be required for the task the following day.

Following the classification, panelists were given the opportunity to discuss their classifications. This discussion was a whole group discussion so that all panelists would hear all comments. Panelists gained a sense of how their classification judgments compared to others in the group. They reported that this discussion helped them feel more confident about their preparation for the task the following day.

THE BOOKLET CLASSIFICATIONS

For the Booklet Classifications, teachers were asked to classify 40 booklets into NAEP achievement level categories using only the achievement level descriptions as the criterion. The data reported in Table 9 show the correspondence between teachers' classification of booklets and the empirical score classifications that were based on MLE estimates of student performances. In this table, the data reported are for classifications by 11 teachers of 40 booklets selected from the empirical score classifications to produce this distribution of booklets: 7 Below Basic, 13 Basic, 13 Proficient, and 7 Advanced booklets.

Table 9
Correspondence of Teachers' Classifications of Student Booklets
into Achievement Level Categories and Empirical Score Classifications
of Student Booklets into Achievement Level Categories

Table N =440	Achievement level classification of student booklets by teachers			
Achievement level classification by empirical scores of student booklets (ACT NAEP-Like Cutscores)	Below Basic (n=137)	Basic (n=155)	Proficient (n=104)	Advanced (n=44)
Below Basic (<149.2) (n=77)	15.9% (n=71)	1.4% (n=6)	0.0 (n=0)	0.0 (n=0)
Basic (149.2 – 165.39) (n=143)	14.3 (n=63)	17.7 (n=78)	0.5 (n=2)	0.0 (n=0)
Proficient (165.4 – 177.89) (n=143)	0.6 (n=3)	16.2 (n=71)	14.1 (n=62)	1.6 (n=7)
Advanced (≥ 177.9) (n=77)	0.0 (n=0)	0.0 (n=0)	9.1 (n=40)	8.4 (n=37)

Bold entries are for cells that would represent "hits."

P_A = .561
P_E = .263
K = .404

Panelists classified 56% of the booklets at the same achievement level as the empirical score classification based on MLE estimates for the booklets. Overall, they tended to classify the booklets at the same level as the empirical score level, or lower. The exact correspondence between teacher judgments and student performances was higher in this classification of booklets than for the previous two classifications involving judgments about achievement and performance of specific students. Classifications were within one achievement level of the empirical score classification for all except three of the booklets. Those three booklets were actually scored within the Proficient level, but they were classified as Below Basic. Only 9 booklets (2%) were judged to represent performance higher than the empirical score classification, and 177 booklets (40%) were judged to represent performance below the empirical score classification.

This pattern was exactly opposite that found when these same teachers classified students in the SCS portion of the study. When teachers classified the expected performance *of their own students* and overall knowledge and skills *of their own students*, they classified student achievement at higher levels than student performance on the special form of NAEP would seem to warrant. When teachers classified actual performances of *anonymous* students on this special form of NAEP, however, their judgments led to classifications that were lower than the students' actual performance on the special form of NAEP would seem to warrant. Based on the consistent

results of the booklet classification studies conducted by ACT, the decision would be to move the NAEP cutscores for these subjects and grades to higher scores.

Table 10 reports the mean, minimum, and maximum ACT NAEP-like score values for performance scores of the booklets classified at each achievement level. These data allow more detailed evaluation of the relative levels of performances judged to represent different achievement levels. With one exception, a logical pattern of average scores was found for the booklets classified at each level. This indicated that teachers were able to discern a relative ordering of performance, even though they generally classified booklets at one level lower than the empirical score classification level. For example, the average score (143) of the booklets that were scored below the Basic cutscore but classified at the Basic level was higher than those scored below the Basic cutscore and classified at the Below Basic level (135). Similarly, the average score of booklets scored within the range of the Basic cutscores and classified at the Proficient level was 160, compared to those classified within the Basic range (159), and the Below Basic range (157). The exception was the average score of the three booklets that were classified in the Below Basic category but actually scored within the Proficient range. Those three booklets had an average score (171) that was about the same as the 62 booklets scored and classified within the Proficient range. At least one of the booklets judged to be in the Below Basic category was scored 174 on the ACT NAEP-like scale—well above the Proficient cutscore of 165.

Table 10
Mean, Minimum, and Maximum ACT NAEP-Like Score Values for
Performance Scores of Student Booklets Classified at Each Achievement Level

Table N =440 Achievement Level Classification of MLE Score Estimates of Student Performance (ACT NAEP-Like Cutscores)	Achievement level classification of student booklets by teachers			
	Below Basic (n=137)	Basic (n=155)	Proficient (n=104)	Advanced (n=44)
Below Basic (<149.2) (n=77)	134.8* (119.3/146.3) (n=71)	143.0 (136.7/146.3) (n=6)	0.0 (n=0)	0.0 (n=0)
Basic (149.2 – 165.39) (n=143)	157.2 (153.6/163.7) (n=63)	158.5 (153.6/163.7) (n=78)	160.0 (158.5/161.6) (n=2)	0.0 (n=0)
Proficient (165.4 – 177.89) (n=143)	170.5 (169.0/173.8) (n=3)	165.4 (169.0/175.6) (n=71)	171.9 (169.0/175.6) (n=62)	173.6 (172.4/175.6) (n=7)
Advanced (≥ 177.9) (n=77)	0.0 (n=0)	0.0 (n=0)	181.6 (179.7/184.7) (n=40)	182.2 (179.9/184.7) (n=37)

***Bold** numbers are the mean ACT NAEP-like score value for the booklets classified at the level. Numbers within parentheses are minimum/maximum values of booklets classified at the level.

COMPARISONS OF EACH TEACHER’S THREE CLASSIFICATIONS OF THEIR OWN STUDENTS

The design for the civics study was changed somewhat from that used in 1995 in order to provide the opportunity to have the same panelists classify student booklets (the BCS) and their own students (the SCS). These study findings indicated that the cutscores were set too high when performance levels on NAEP were compared to teachers’ estimations of the achievement of their

students relative to the performance criteria. These study findings also indicated that the cutscores were set too low when performance levels on NAEP were compared to teachers' evaluations of students' actual performance on the assessment. There was a lack of certainty regarding the generality of these findings, however, because the differences could result from judgments by two different sets of panelists. Findings from this study have revealed the *same patterns of findings* produced in the two different studies for each subject in the 1994 ALS process. This adds credibility to the findings of both sets of studies.

A final set of comparisons for a limited number of students was possible, and these comparisons provided an even more direct test of whether these same patterns hold. The expected patterns were that teachers' classifications of their own students would be higher than the students' empirical performance relative to the NAEP achievement levels; and teachers' classifications of the performance of students whose identity was unknown to them would be lower than the students' empirical performance. That is, the same pattern found for teachers' classifications, overall, was expected for the classifications of the teachers' own students.

Thirty-seven of the 40 student booklets included in the BCS for civics were assessments of students in classrooms taught by the nine teachers in the study. During the BCS portion of the study, only one teacher commented that she could match the identity of one student to the booklet. Whether more teachers discerned the identity of the examinees is not known. It seems unlikely that this was a frequent occurrence since no one else commented on this during the study.

The data in Table 11 show the patterns found across the three classifications. Consistently, the data in Table 11 show that these teachers *tended* to classify the overall knowledge and skills and the expected performance on the special NAEP of their own students at the same level or at a higher achievement level than the level at which the students performed on the special NAEP. When the teachers read the responses of those same students, in order to classify the performance represented in the booklet, they *tended* to classify the booklet at a level lower than the level at which the student performed. Furthermore, teachers classified the examinee booklets of their students at lower levels than they classified the expected performance by the same students. Finally, teachers classified the examinee booklets of their students at lower levels of achievement than they classified the overall level of knowledge and skills for the same students.

It is particularly interesting to examine the correspondence between teachers' classifications of the test booklets of their students, i.e., how teachers classified the actual performance of their students, and their classifications of the expected performance of the same students on an assessment with the same attributes of the special form of NAEP that was administered to their students. The data in the lower right section of Table 11 show how teachers of the students in the study classified these two factors. Expectations of how the student would perform on such an assessment corresponded to the teacher's evaluation of the student's performance for only 16 of the 37 students (43%) for which all comparable data were available. Teachers classified the actual performance, i.e., the booklets, of 19 (51%) of the students at a lower level than the level at which the expected performance was classified. When booklet classifications were compared to the teacher's classification of the same student's overall knowledge and skills, the findings again showed that teachers classified the booklets of 20 students (54%) at a lower level than they classified the overall civics knowledge and skills of the same students. Those data are reported in the middle section of the right side of Table 11. Teachers classified the performance of only 15 students (40%) at the same level as the overall knowledge and skills classification. In all, booklets for only two students were classified at a higher level of achievement than that judged to represent the student's overall knowledge and skill in the subject of the assessment.

Table 11
Relationships Between and Among Empirical Score Performance Classifications
and Teacher Judgment Classifications of Performance by NAEP Achievement Levels

		Overall Knowledge & Skills Level				Expected NAEP Performance Level				Booklet Classification			
		BB	B	P	A	BB	B	P	A	BB	B	P	A
Empirical Classification	BB (6)	3	3	0	0	5	1	0	0	5	1	0	0
	B (12)	1	6	5	0	1	6	5	0	7	4	1	0
	P (12)	0	2	4	6	0	3	6	3	0	7	5	0
	A (7)	0	0	1	6	0	0	2	5	0	0	6	1
	Total	4	11	10	12	6	10	13	8	12	12	12	1
Overall Knowledge & Skills	BB (4)					4	0	0	0	5	1	0	0
	B (11)					2	9	0	0	7	4	1	0
	P (10)					0	1	9	0	0	7	5	0
	A (12)					0	0	4	8	0	0	6	1
	Total					6	10	13	8	12	12	12	1
Expected Performance	BB (6)									5	1	0	0
	B (10)									5	4	1	0
	P (13)									2	5	6	0
	A (8)									0	2	5	1
	Total									12	12	12	1

PANELISTS' EVALUATIONS OF THE PROCEDURES AND INFLUENCES ON THEIR CLASSIFICATIONS

Throughout the three-day process, panelists were asked for their comments and reactions to research procedures in an effort to examine their perceptions of the classification process. Teachers completed four evaluation questionnaires similar to those administered during the ALS process. Each form encouraged panelists to comment on any aspect of their experience that was not covered specifically on the questionnaires, in addition to many aspects that were covered on the questionnaires.

Overall, the panelists indicated that they were generally confident and satisfied with their understanding of the classification tasks and the procedures for performing them. There were no apparent problems or areas for concern based on the responses to the evaluations. Many commented on their enjoyment of the research experience and expressed appreciation for the opportunity to participate in the study.

RESPONSES TO QUESTIONS REGARDING FUNDAMENTAL FEATURES OF THE CLASSIFICATION PROCEDURES

Data reported in Table 12 show the average responses (5=most positive and 1=most negative) to questions about the SCS and BCS classification processes. The first questionnaire administered referred to classifications of the overall civics knowledge and skills of the students. The second questionnaire referred to the "second" classification. Panelists were instructed to understand that this second classification was of their estimates of their students' performance on an assessment with the attributes of the special form of Civics NAEP, relative to the achievement levels descriptions. Similar questions were asked on the questionnaire following the booklet classification procedure. Civics teachers indicated that they were very clear in their understanding of the task and the descriptions of student performance within the three achievement levels. All mean scores for these questions were ≥ 4.55 (5 = *absolutely clear*). Panelists' responses were even more positive when asked about their understanding of the instructions for what they were to do during the two classification sessions. When reviewing the responses to this question for both classifications, 19 of the 22 combined responses were *absolutely clear*. The amount of time allowed to complete the classifications was reported as generally about right (= 3), or somewhat longer than needed.

Responses were somewhat less positive for the BCS than for the SCS when panelists were asked about their understanding of the ALDs relative to the ease of identifying student booklets within a given achievement level. This was particularly true for the practice set of booklets. Panelists' responses reflected an increase in understanding of the booklet classification process as evidenced by the increase in all of the mean scores from the initial practice round to the actual classification session. These data are supported by teachers' open-ended comments that indicated that the discussions during the practice session clarified their concept of the booklet classification task. As with the SCS, panelists' responses were quite positive when asked about the clarity of the instructions for what they were to do during the actual booklet classification session. Ten of 11 responses were *absolutely clear*. Teachers reported their level of understanding of the task they were to do during the booklet classification session as even higher than that for the SCS. The amount of time allowed to complete the classifications was reported as generally about right. One teacher, however, reported feeling rushed and distracted when others were turning in their finished classifications while s/he was still working.

It is noteworthy that most of the teachers thought that the civics knowledge and skills described by the ALDs were similar to the criteria they use to evaluate their students. Only 2 of the 11 panelists indicated that the ALDs were less than *somewhat similar* to their own evaluation criteria.

Table 12
Mean Responses to Questions About Fundamental Aspects of the Classification Process

Question	SCS Mean	Question	BCS Mean
<i>Understanding of Basic</i> (5 = Absolutely Clear; 3 = Somewhat Clear; 1 = Not at All Clear)			
When I classified the level of civics achievement of my students (SCS1), my understanding of the description of student performance within the <u>Basic level</u> of achievement was:	4.55	When I classified the practice set of booklets, my understanding of the description of student performance at the <u>Basic level</u> of achievement made it easy to identify student performance within the Basic level.	3.91
The second time I classified the level of civics achievement of my students (SCS2), my understanding of the description of student performance within the <u>Basic level</u> of achievement was:	4.64	When I classified the booklets, my understanding of the description of student performance at the <u>Basic level</u> of achievement made it easy to identify student performance within the Basic level.	4.45
<i>Understanding of Proficient</i> (5=Absolutely Clear; 3=Somewhat Clear; 1=Not at All Clear)			
When I classified the level of civics achievement of my students, my understanding of the description of student performance within the <u>Proficient level</u> of achievement was:	4.55	When I classified the practice set of booklets, my understanding of the description of student performance at the <u>Proficient level</u> of achievement made it easy to identify student performance within the Proficient level.	3.82
<i>Understanding of Advanced</i> (5=Absolutely Clear; 3=Somewhat Clear; 1=Not at All Clear)			
When I classified the level of civics achievement of my students, my understanding of the description of student performance within the <u>Advanced level</u> of achievement was:	4.50	When I classified the practice set of booklets, my understanding of the description of student performance at the <u>Advanced level</u> of achievement made it easy to identify student performance within the Advanced level.	4.27
<i>Instructions</i> (5= Absolutely Clear; 3=Somewhat Clear; 1=Not at All Clear)			
The instructions on what I was to do during this classification session were:	4.82	The instructions on what I was to do during this (practice) classification session were:	4.63
The instructions on what I was to do during the second classification session were:	4.91	The instructions on what I was to do during the booklet classification session were:	4.91
<i>Level of Understanding</i> (5 = Totally Adequate; 3 = Marginally Adequate; 1 = Totally Inadequate)			
My level of understanding of the task I was to perform during this classification session was:	4.73	My level of understanding of the task I was to perform during this classification session was:	4.64
My level of understanding of the task I was to perform during the second classification session was:	4.73	My level of understanding of the task I was to perform during the booklet classification session was:	4.82

Question	SCS Mean	Question	BCS Mean
Level of Confidence (5 = Totally Confident; 3 = Somewhat Confident; 1 = Not at all Confident)			
The most accurate description of my level of confidence in the estimates I made during this session of achievement levels categories of my students is that I was:	3.82	The most accurate description of my level of confidence in the estimates of achievement levels categories for the practice booklets during this session is that I was:	3.91
The most accurate description of my level of confidence in the estimates I made during this (second) session of achievement levels categories of my students is that I was:	3.55	The most accurate description of my level of confidence in my classifications of booklets is that I was:	3.91
Amount of Time (5 = Far Too Long; 3 = About Right; 1 = Far Too Short)			
The amount of time allowed to complete the first classification session was:	4.00	The amount of time allowed to complete the practice classification session was:	2.91
The amount of time allowed to complete the second classification session was:	2.91	The amount of time allowed to complete the classification session was:	3.18

Responses were somewhat less positive to questions about panelists' level of confidence in their estimates of achievement levels categories for their students. Mean scores for confidence dropped to ≥ 3.55 for the teachers' estimates of students' expected performance on the special form of NAEP. The open-ended responses to the questions about teachers' confidence indicated that many teachers felt very confident in estimating their students' general knowledge and skills in civics for the first classification of the SCS (mean = 3.82). For the second classification, however, several teachers expressed less confidence in estimating their students' performance on the Civics NAEP because of many unpredictable factors, such as student motivation. A review of the frequencies of responses to the questions about confidence revealed that only one panelist indicated a lack of confidence (either 1 or 2) and 10 responded positively (either 3, 4 or 5) for each classification.

Level of Confidence in Estimates

Mean	1 (Not at All)	2	3 (Somewhat)	4	5 (Totally)
Civics Gen 3.82	0	1 (9.1%)	2 (18.2%)	6 (54.5%)	2 (18.2%)
Civics NAEP 3.55	1 (9.1%)	0	3 (27.3%)	6 (54.5%)	1 (9.1%)

As with the SCS, responses for the BCS were relatively less positive to questions about panelists' level of confidence in their estimates of achievement level categories for student booklets. After participating in the practice classification session, however, all panelists responded positively when asked if they felt confident that they would be able to judge performance of students relative to the ALDs.

After participating in the practice classification session, I feel confident that I will be able to judge performances of students relative to the ALDs

Mean	1 (Totally Disagree)	2	3 (Somewhat)	4	5 (Totally Agree)
4.36	0	0	1 (9.1%)	6 (54.5%)	4 (36.4%)

To probe the confidence issue further, panelists were asked about the probable accuracy of their classifications of students' general knowledge and skills in civics according to the achievement levels descriptions. Eight panelists thought they *correctly estimated* what their students know and can do in civics, 2 thought they *somewhat under estimated*, and only 1 thought s/he *somewhat over estimated*. Results of the classification of overall civics knowledge and skills indicated, however, that teachers tended to over estimate the performance.

Classifications of General Civics Knowledge Are Likely To Be

Mean	1 (Under Estimated)	2	3 (Correctly Estimated)	4	5 (Over Estimated)
2.91	0	2 (18.2%)	8 (72.7%)	1 (9.1%)	0

Similarly, after discussing their practice classifications of booklets and thinking about the ALDs, panelists were asked about their tendency to classify booklets too high or too low before they began the actual booklet classification session. Seven teachers thought they tended to classify booklets *about right*, and 4 thought they tended to classify booklets either *somewhat* or *far higher*. Interestingly, *none* thought they tended to classify booklets lower. The BCS findings, of course, show just the opposite. These teachers tended to classify booklets at the same level, or one level lower than the empirical level.

After discussing your practice classifications and thinking about the ALDs, do you think you tended to classify booklets too high or too low? I tend to classify booklets:

Mean	1 (Far Lower)	2	3 (About Right)	4	5 (Far Higher)
3.45	0	0	7 (63.6%)	3 (27.3%)	1 (9.1%)

IDENTIFICATION OF FACTORS THAT INFLUENCED PANELISTS' JUDGMENTS

It is of interest to examine the factors that influence panelists' judgments when providing classifications of student achievement and expected performance. Data reported in Table 13 show the mean scores and percentages of panelists' responses when asked about various factors that could have influenced their judgments. It is important to remember that although the questions pertain to similar or identical issues, the questions are asked at different stages of the process and relate to different tasks. Responses labeled as *Civ Gen* were collected on the first process evaluation questionnaire. That questionnaire was administered after the classification of students' overall civics knowledge and skills. *Civ NAEP* responses are to questions on the second process evaluation questionnaire that addressed teachers' perceptions when classifying their students' expected performance on the special form of the Civics NAEP administered for the study.

Table 13
Factors That Influenced Teachers' Judgments for
Student Classifications as Reported After Classification of Overall Civics Knowledge and Skills and
Classification of Expected Performance on Special Form of Civics NAEP

Question	Subject	Mean Response	Percentages				
			1 Very Little	2	3 Moderately	4	5 Very Much
Performance in Other Courses (5=Very Much; 3=Moderately; 1=Very Little)							
When I provided this set of student classifications, my judgments were influenced by my knowledge of the students' academic performance in other courses.	Civ Gen (n=11)	2.09	54.5	0.0	36.4	0.0	9.1
	Geo (n=19)	1.58	73.7	10.5	5.3	5.3	5.3
	Hist (n=16)	2.06	37.5	37.5	12.5	6.3	6.3
When I provided the second set of student classifications, my judgments were influenced by the students' academic performance in other courses.	Civ NAEP	1.55	72.7	0.0	27.3	0.0	0.0
	Geo	1.72	55.6	22.2	16.7	5.6	0.0
	Hist	1.19	81.3	18.8	0.0	0.0	0.0
Overall Civics Knowledge and Skills (5=Very Much; 3=Moderately; 1=Very Little)							
When I provided the second set of student classifications, my judgments were influenced by my first classifications.	Civ NAEP	2.82	9.1	18.2	54.5	18.2	0.0
	Geo	3.61	0.0	5.6	50.0	22.2	22.2
	Hist	3.13	6.3	12.5	56.3	12.5	12.5
Grade in Course (5=Very Much; 3=Moderately; 1=Very Little)							
When I provided the second set of student classifications, my judgments were influenced by the students' grade in my course.	Civ NAEP	2.27	36.4	18.2	27.3	18.2	0.0
	Geo	2.89	11.1	11.1	55.6	22.2	0.0
	Hist	2.19	43.8	18.8	25.0	0.0	12.5
Items on the Civics NAEP (5=Very Much; 3=Moderately; 1=Very Little)							
When I provided the second set of student classifications, my judgments were influenced by the items included on the grade 8 NAEP.	Civ NAEP	4.36	0.0	0.0	18.2	27.3	54.5
	Geo	3.50	5.6	11.1	27.8	38.9	16.7
	Hist	4.50	0.0	0.0	6.3	37.5	56.3
Test Length (5=Very Much; 3=Moderately; 1=Very Little)							
When I provided the second set of student classifications, my judgments were influenced by the adverse effect the length of the special NAEP assessment might have on student performance.	Civ NAEP	2.45	27.3	18.2	36.4	18.2	0.0
	Geo	3.22	5.6	22.2	33.3	22.2	16.7
	Hist	3.13	6.3	18.8	37.5	31.3	6.3

Question	Subject	Mean Response	Percentages				
			1 Very Little	2	3 Moderately	4	5 Very Much
Motivation (5=Very Much; 3=Moderately; 1=Very Little)							
When I provided the second set of student classifications, my judgments were influenced by the adverse effect on student motivation that might result from the fact that student scores will be reported to no one.	Civ NAEP	3.18	18.2	9.1	18.2	45.5	9.1
	Geo	2.50	27.8	27.8	16.7	22.2	5.6
	Hist	3.19	6.3	6.3	56.3	25.0	6.3

Influences Reported in Second Classification of Overall Civics Knowledge and Skills

As noted previously, an on site decision was made to create another opportunity for teachers to classify the overall civics knowledge and skills of their students and to indicate what factors influenced their judgments. After the first SCS classification, teachers were again asked to categorize the civics knowledge and skills of their students relative to the ALDs. This time they had only half the number of students to classify. The classification was the same as for the first session except a list of six factors was included on the classification form. Panelists were asked to rate the influence of each factor on their classification of each student (5=very large influence and 1=no influence). The factors were:

- Overall knowledge and skills in all subjects
- Overall knowledge and skills in civics
- Test-taking behavior
- Achievement levels descriptions
- Items on the Civics NAEP
- Grade(s) in my course

Many of these same factors had already been included on the subsequent process evaluation questionnaire distributed after teachers estimated their students' performance on the special form of the Civics NAEP. It is uncertain how teachers' second classifications were affected by the introduction of these factors on Form 1B, or how teachers' responses to the questionnaire items were affected. To examine this issue, data for the SCS for geography and U.S. history are included in Table 13. The pattern of responses does not appear to be unusual when comparing the responses to the second questionnaire in the civics study with similar questions asked at parallel junctures in the studies for geography and U.S. history. Therefore, it would appear that the factors introduced on Form 1B had little affect on teachers' responses on the second civics process evaluation questionnaire. Whether the responses reflect the actual impact on classifications is not known.

Influence of Overall Knowledge and Skills in All Subjects

When teachers were first asked about *the influence of students' knowledge and skills in other courses* on their classification of students' knowledge and skills in civics, most teachers responded that the influence was *moderate to very little* (Civ Gen mean = 2.09). The same general pattern of responses emerged when teachers estimated student performance on the special form of the Civics NAEP (Civ NAEP mean = 1.55). These responses were similar to those found

for geography and U.S. history in that performance in other courses did not influence classifications “very much.”

Influence of Overall Knowledge and Skills in Civics

Teachers reported that their students’ general civics knowledge *moderately* influenced their estimates of students’ expected performance on the special version of the Civics NAEP.

Influence of Students’ Grade in Teachers’ Course

Teachers’ responses were spread across all but the most positive response when reporting the influence of students’ grades in the teachers’ course when classifying their students’ expected performance on the special form of the Civics NAEP. The modal response was *very little* and the mean reflected a relatively low level of influence (Civ NAEP mean = 2.27).

Influence of Items on the Civics NAEP

As would be expected, 100% of teachers reported that they were influenced by the items on the Civics NAEP at least moderately, when estimating their students’ expected performance on the special form of the Civics NAEP. More than half of the teachers responded that they were *very much influenced* by this factor.

Influence of Length of Special NAEP and Motivation

Teachers indicated that the length of the special form of the Civics NAEP did not greatly influence them in forming their judgments of how their students would perform on the assessment. Student motivation seemed to be a greater influence (mean = 3.18) on teachers’ estimates of their students’ performance on the Civics NAEP than assessment length (mean = 2.45).

COMMENTS AND OPEN-ENDED RESPONSES ABOUT FACTORS INFLUENCING CLASSIFICATIONS

Process Evaluation Questionnaires

On the open-ended responses of the process evaluation questionnaires, teachers commented that they also considered other influences not targeted specifically on the forms. For the classifications regarding overall civics knowledge and skills and expected performance on the special NAEP, such factors included the civics content they taught to their students, limitations of special needs students, and students’ reading and language skills.

For the booklet classifications, factors that most frequently appeared in teachers’ comments when asked what they considered were the ALDs, the correctness and completeness of constructed responses, and the difficulty level of multiple choice items. A few teachers indicated that they were influenced by whether or not students attempted to answer constructed response items. Many teachers commented on the importance they placed on the quality of students’ constructed responses and the need for students to produce a full and detailed answer to receive full credit. Teachers reported that they perceived student performance to be lower when answers to constructed response items lacked details. These comments provided further evidence of the fact

that panelists use noncompensatory decision strategies when making holistic judgments of student performances⁷.

SCS2 and BCS Classification Forms. Teachers were asked to comment on the types of things they took into consideration when classifying their students' expected performance on the Civics NAEP (SCS2). After classifying 40 anonymous assessment booklets (BCS), they were asked to comment on the types of things they took into consideration when classifying a collection of 5 booklets that had been pre-specified for comments. These booklets selected for teachers' comments were assigned in such a way that 9 teachers commented on at least one of their own student's booklet. Of particular interest are the comments from the same teacher for the same student for both the SCS2 and the BCS. Table 14 presents this information. In general, teachers' comments about the expected performance of their individual student tended to be somewhat personal and related to a variety of characteristics of the student. In contrast, comments about the performances represented in the booklets were focused on how well the student satisfied the standard described by the ALDs. The teachers' comments suggested that factors that influence teachers' judgments of performance for a known, individual student can be quite different from the factors teachers consider when forming their judgment of performance for anonymous students.

⁷ See the report by Bay (2000) for a description of a study by ACT to examine decision strategies by panelists in making holistic judgments.

Table 14

**Teachers' Comments About Factors That Influence Estimates of Their Own Students' Performance for
Expected Student Performance on the Special Form of the Civics NAEP (SCS 2) and
Classifications of NAEP Booklets (BCS)**

I.D.	SCS 2 Comments	SCS 2	BCS 2	Empirical	BCS 2 Comments
Teacher 5 Booklet 5	Not a strong student, severe emotional and interpersonal problems, often absent, a retainee	Low BB	BB	BB	Many gaps in basic knowledge, MC questions about basic civics and constitution were wrong, seemed to be little understanding (necessary in many criteria for basic), most CR answers showed little understanding of basic ideas and facts
Teacher 7 Booklet 6	Borderline student	High BB	BB	B	Does not understand 14 th amendment, functions of the 3 branches of government, function of local government, rights of citizenship, U.S. involvement with foreign affairs, labor unions, supreme court actions and cases, civic and civil organizations, strengths and weaknesses of Articles of Confederation, due process, some trouble with pictures, graphs...
Teacher 9 Booklet 7	Special magnet program, reading and work study skills, sporadic motivation	High P	P	A	MC answers demonstrate his proficiency in concepts presented – separation of powers, Bill of Rights and checks/balances. Some CR lean toward Advanced, but are not totally detailed to show application.
Teacher 12 Booklet 9	Achievement levels, test item pool, knowledge base, motivation, competitive, high expectations of herself	High A	P	A	A lot of the short response answers show insight, but one or two were off target. The student did not have a complete understanding of the documents or foreign relations – the two areas they missed multiple choice questions. Their understanding fell in the Proficient range.
Teacher 8 Booklet 13	Motivated and good content knowledge, reading, vocabulary and good thinker/writer	High P	P	P	Student showed pretty solid grasp of content knowledge, attempted constructed responses were answered accurately for the most part, but some were only partially answered, not a lot of elaboration, but got the gist of most; able to interpret non-text based information visual graphs, charts, cartoons
Teacher 13 Booklet 17	Does not communicate in class, ability level	Mid B	B	P	I look at the kinds of answers that were given, example question 13, the student did not answer the question and some were incomplete
Teacher 6 Booklet 29	Low motivation but very high skill level and test taker	High A	B	P	The CR showed no real ability to explain the how and why necessary to get full credit for them. The MC showed good knowledge of basic civics facts. This one was tough because it did show knowledge and the student was able to interpret charts and graphs, but the CR just did not show an ability to do any real thinking with the basic knowledge

I.D.	SCS 2 Comments	SCS 2	BCS 2	Empirical	BCS 2 Comments
Teacher 7 Booklet 32	Did not try in class	Low BB	BB	BB	Restated answer #4 for #5. Did not understand judicial review. Did not understand separation of local, state, national government. Could not read charts, graphs..., did not know purpose of labor unions, did not know function of the supreme court, did not know strengths and weaknesses of Articles of Confederation, lacks knowledge of due process, cannot distinguish between foreign and domestic issues, does not know functions of the branches of government, does not know purpose of bill of rights, Federalist Papers, rights of citizenship, cannot distinguish responsibilities of local and national government
Teacher 2 Booklet 36	Solid knowledge base, able to think and learn abstractly	High P	B	P	Student was unable to fully identify functions of elections, checks/balances, interest groups. Student was unable to fully identify branches/functions of federal government
Teacher 10 Booklet 40	7 th grader – little civics instruction, average test scores, average grades, on level reading	Mid B	BB	B	Missed far too many MC questions, most of the answers to the CR questions were weak at best, left out too many questions including the chart on the checks and balances of the different branches of government

CONCLUSIONS

The general rationale behind this research design is that if teachers who participated in the five-day achievement levels-setting process *cannot* use the descriptions to judge student performance, then it is unlikely that anyone can. So, if their classifications were wildly *different* from the performances in the booklets, we would tend to think that the cutpoints do not denote performance consistent with the ALDs.

Findings show patterns similar to those found in 1995. That is, teacher panelists tend to classify their own students higher than their performance levels on the special form of NAEP developed for the study. This is true for classifications of the overall civics knowledge and skills of their students and for classifications of expected performance on the special form of the Civics NAEP developed for this study. The highest classification levels were for students' overall knowledge and skills in Civics, followed by classifications of students' expected performance on the special form of the Civics NAEP. Both of these classifications tended to be at or above the empirical score classification of the student's performance. When the same teachers were asked to classify the performance of the students, represented in the special Civics NAEP test booklets, they tended to classify those at or below the empirical score classification of the student's performance.

DISCUSSION OF FINDINGS

This is the first booklet classification study conducted with reliable measures at the individual student level. Previous studies have indicated that the plausible values tend to increase performance levels, relative to the raw score classifications. This is ACT's first NAEP study with a BCS design for which the scores were not derived through plausible values. The findings of this study concur with the previous findings by ACT regarding standard setting with a booklet classification method. That is, the standards set with a booklet classification method will be higher than those set with the item-by-item method used for the NAEP ALS process.

Further, even when teachers are well-trained in the NAEP achievement levels descriptions and are familiar with the NAEP assessment pool, they are still likely to overestimate the knowledge and skills of their students, when those estimates are compared to the students' actual performance on the NAEP. The findings of this study and other conducted by ACT for the NAEP ALS process are consistent with other studies regarding teachers' abilities to judge student performances on non-classroom test instruments. Other research related to teachers' abilities to judge the performance of their students indicated that teachers are only moderately proficient at estimating the *relative* level of their students' performance on standardized test (Hoge & Coladarci, 1989). That is, teachers can do fairly well in ordering their students' performance on standardized tests. Teachers do less well, however, when it comes to their estimating the *actual* level of their students' achievement relative to the students' standardized test scores. Teachers consistently tend to overestimate their students' standardized test performance (Perry & Meisels, 1996).

The study involved only grade 8 teacher panelists, so one cannot be certain that the same results would hold for teacher panelists at other grades. There is no reason to expect that the results for the grade 8 teachers would differ significantly from those for teachers at other grades, but we have no data to test that.

Further, we cannot generalize to a larger population with these results. The study was designed to help focus on the ALS process and the ability of panelists to make rational and reasonable judgments. If the results indicate that these teachers would make significantly different judgments when using the ALDs with respect to expectations about their own students than they made in the rating process, then we would doubt the results of the ALS process. We would doubt that anyone could accurately interpret performance relative to the ALDs. On the other hand, if the results indicate that these teachers made judgments that are very similar to the judgments they made when setting the achievement levels, then we *cannot* say that others would make similar judgments.

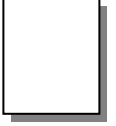
One important feature of this study is that it combined both the SCS and the BCS designs. The 1995 studies in geography and US history showed that teachers tended to classify their own students at a higher level than the students' performance would warrant. We also found that panelists classified booklets of students, unknown to them, at a level lower than the empirical score classification. There appeared to be a rather compelling logic to these patterns, but it was not clear that the findings were a result of differences in the sets of panelists or a more general behavioral judgment finding. Having the same panelists participate in both parts of the study provides the control on panelists to make direct comparisons across different classification tasks. Having teachers' classifications for some students of both the student's overall civics knowledge and skills and expected performance on the special form of the NAEP *and* of their actual performance in the BCS provides data on all four classifications by one teacher for one student.

The results of this study provide information needed to confirm that the general achievement levels-setting process for the 1998 Civics NAEP appeared to "work." That is, panelists were able to use the ALDs in a different setting and for different purposes, and the translations with respect to the score scale seem reasonably on target.

REFERENCES

- ACT (1997). *Setting Achievement Levels on the 1996 NAEP in Science: Final Report, Volume IV: Validity Evidence and Special Studies*. Iowa City, IA: Author.
- ACT (1995a). *Results of the 1994 Geography NAEP Achievement Levels-Setting Pilot Study*. Iowa City, IA: Author.
- ACT (1995b). *Results of the 1994 US History NAEP Achievement Levels-Setting Pilot Study*. Iowa City, IA: Author.
- Bay, L. (2000). "Setting Achievement Levels on the 1998 National Assessment of Educational Progress in Writing: Performance Profiles Study" in Loomis, S.C. (Ed.). *Developing achievement levels on the 1998 National Assessment of Educational Progress in Writing: Research studies*. Iowa City, IA: ACT.
- Carlson, J. (1995). *Estimation of reliability of NAEP IRT proficiency score estimates*. Technical Memorandum, ETS.
- Hanson, B.A. & Bay, L. (1999). *Classifying Student Performance as a Method for Setting Achievement Levels for NAEP Writing*. Paper presented at the annual meeting of the national Council on Measurement in Education, Montreal.
- Hoge, R. D., & Coladarci, T. (1989). Teacher-Based Judgments of Academic Achievement: A Review of Literature. *Review of Educational Research*, 59, 297-313.
- Loomis, S.C. (2000). *Feedback in the NAEP Achievement Levels Setting Process*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.
- National Academy of Education (1993). *Setting Performance Standards for Student Achievement*. Stanford, CA: Author.
- Perry, N. E., & Meisels, S. J. (1996). *How Accurate Are Teacher Judgments of Students' Academic Performance?* (Working paper No. 96-08). Washington, DC: Office of Educational Research and Improvement, U.S. Department of Education.
- Shepard, L.A. (1995). *Implications for Standard Setting of the National Academy of Education Evaluation of the National Assessment of Educational Progress Achievement Levels*. In National Assessment Governing Board and National Center for Education Statistics, *Joint Conference on Standard Setting for Large-Scale Assessments: Proceedings*, Volume II, Washington, D.C.: U.S. Government Printing Office.

Appendix



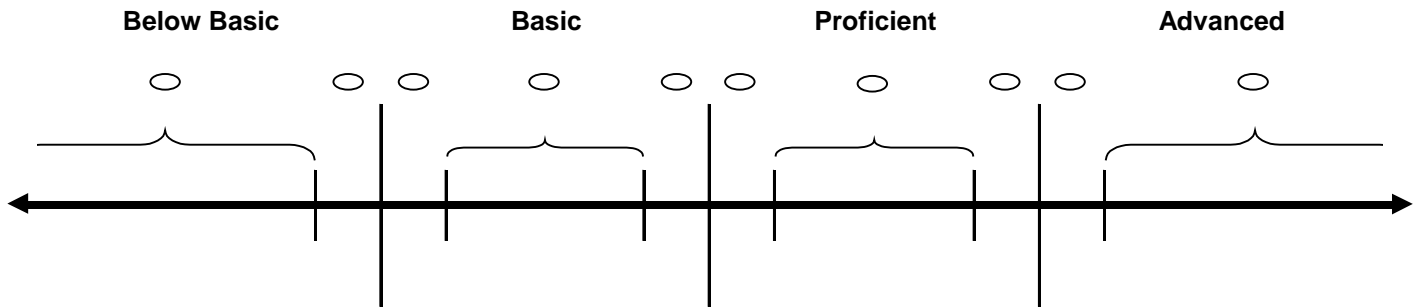
Similarities Classification Study

Classification 1

«Student_Name»
Student

«Teacher_Name»
Teacher

Cannot classify (please explain): _____



My level of confidence regarding this achievement level classification is (mark only one):

- High
- Medium
- Low

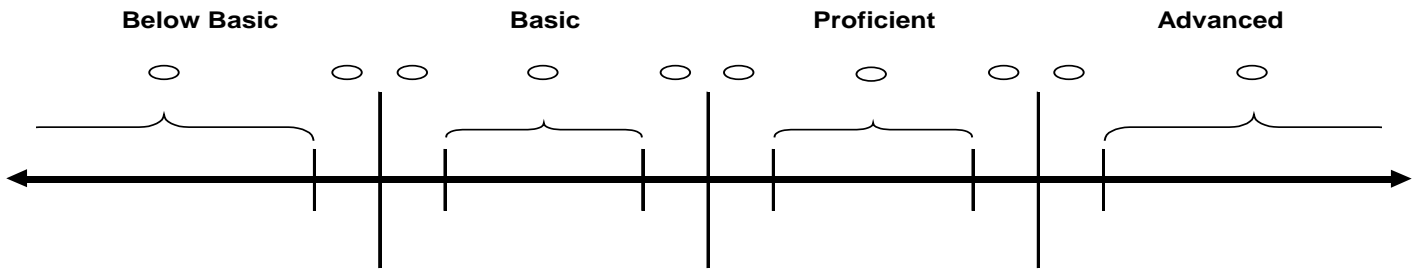
Similarities Classification Study

Classification 1b

«Student_Name» _____
Student

«Teacher_Name» _____
Teacher

Cannot classify (please explain): _____



My level of confidence regarding this achievement level classification is (mark only one):

- High
- Medium
- Low

Here is a list of factors that could have influenced your classification of this student.
Please rate each factor.

	Very Large Influence 5	4	Some Influence 3	2	No Influence 1
Overall knowledge and skills in all subjects					
Overall knowledge and skills in civics					
Test-taking behavior					
Achievement levels descriptions					
Items on the Civics NAEP					
Grade(s) in my course					

Similarities Classification Study

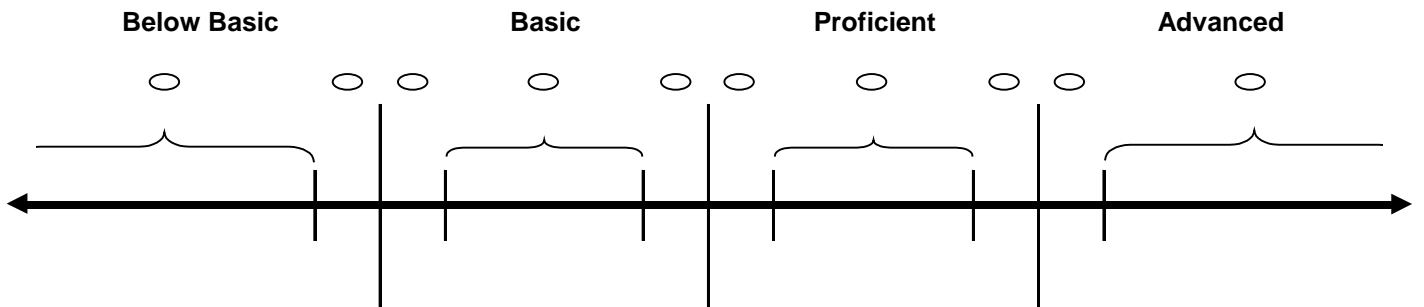
Classification 2

«Student_Name»

Student

«Teacher_Name»

Teacher



My level of confidence regarding this achievement level classification is (mark only one):

- High
- Medium
- Low

Please comment on the types of things you took into consideration when classifying this student:

Booklet Classification Form

Below Basic		Basic		Proficient	Advanced	Below Basic		Basic		Proficient	Advanced
1.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	21.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>		
2.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	22.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>		
3.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	23.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>		
4.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	24.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>		
5.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	25.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>		
6.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	26.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>		
7.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	27.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>		
8.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	28.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>		
9.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	29.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>		
10.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	30.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>		
11.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	31.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>		
12.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	32.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>		
13.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	33.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>		
14.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	34.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>		
15.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	35.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>		
16.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	36.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>		
17.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	37.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>		
18.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	38.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>		
19.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	39.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>		
20.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	40.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>		

Validation Study Participation

**Civics NAEP Validation Research Study
Achievement Levels-Setting Process
Grade 8 Teacher Panelists**

**Ritz-Carlton Hotel, St. Louis
July 9-11, 1999**

Agenda

Thursday, July 8

5:00 - 5:30 p.m. *Amphitheater Prefunction*
Check-in at ACT Registration Desk, 2nd Floor. Get name tag, final agenda, and information about transportation downtown tonight.

Friday, July 9

8:00 a.m. *Consulate*
Continental Breakfast

8:30 a.m. Welcome: Susan Loomis (ACT) and Mary Lyn Bourque (NAGB)
General Orientation Session: Susan Loomis

- Overview of the study
- Review of ALS process

10:00 a.m. Break

10:15 a.m. Review of Framework and Achievement Levels Descriptions (ALDs):
John Patrick

11:15 a.m. Exercises and Discussions to Re-train in Framework and ALDs

- Review Eighth Grade Exemplar items

Noon *Directors*
Lunch

1:00 p.m. Continue Re-training in Framework and Achievement Levels
Descriptions *via* Exercises to practice use of ALDs

1:30 p.m. Instructions for Estimating Student Civics Achievement: Classification
#1

2:00 p.m. Break

2:15 p.m. Student Civics Achievement Classifications #1
Questionnaire #1

4:30 p.m.* Adjournment

* This is an approximate time only. The actual amount of time required will depend upon the number of your students in the study.

Saturday, July 10

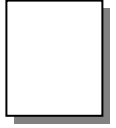
8:00 a.m.	<i>Consulate</i> Continental Breakfast
8:30 a.m.	Review Achievement Levels Descriptions and Framework <i>via</i> Discussions and Exercises <ul style="list-style-type: none">• Student booklets• Grade 8 item pool
10:00 a.m.	Break
10:15 a.m.	Instructions for Estimating Student Civics Achievement: Classification #2
10:30 a.m.	Student Civics Achievement Classifications #2 * Questionnaire #2
Noon - 1:30 p.m.	<i>Promenade</i> Lunch Buffet
2:30 p.m.	The Booklet Classification Study <ul style="list-style-type: none">• Instructions in Booklet Classification Process
2:45 p.m.	Practice Classification Session
3:45 a.m.	Break
4:00 p.m.	Discussion of Practice Classifications
4:45 p.m.	Questionnaire #3
5:00 p.m.	Adjourn

* You may break for lunch before you finish, if necessary. It is, of course, best to complete your classifications without a long break.

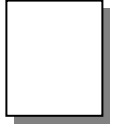
Sunday, July 11

- 8:30 a.m. *Consulate*
Continental Breakfast
- 9:00 a.m. Booklet Classifications
- Information about booklets included in the study
 - Instructions about classification procedures
 - Marking classification forms
 - Review Achievement Levels Descriptions
- 9:30 a.m. Classify Booklets (breaks as needed)
- Noon *Colonnade*
Lunch
- 1:00 p.m. Re-calibration for Continuation of Booklet Classification Study
- Questionnaire #4
- 4:00 p.m. Wrap-Up Session: Questions and Answers
- 4:15 p.m. Adjournment

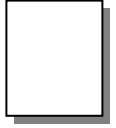
Appendix



Appendix



Appendix



Appendix

