ALPINE TESTING SOLUTIONS

Nebraska Department of Education Statewide Writing Assessment – Standard Setting Report

May 8, 2008

Chad W. Buckendahl, Ph.D.
Brian J. Adams

Acknowledgments

We would like to acknowledge our appreciation to several people who assisted us with this Standard Setting Workshop. Marilou Jasnoch, director of the Statewide Writing assessment, and Jackie Naber at the Nebraska Department of Education were very helpful in organizing this workshop. Marilou also selected the anchor papers and provided content assistance in conducting the standard setting workshop.

We also want to thank the educators whose recommendations contributed to the outcome of the standard setting workshop. Educators who participated in the workshop provided classifications of student performance that resulted in the recommended cut scores. We appreciate their efforts to make the study a success.

Nebraska Department of Education

2008 Statewide Writing Assessment: Grades 4, 8, and 11

Standard Setting Study

Final Report

The purpose of this report is to document the procedures and analyses undertaken to recommend performance standards for the Nebraska Department of Education's *Statewide Writing Assessment* administered in grades 4, 8, and 11. The report summarizes the procedures and the results of the standard setting studies and provides recommendations for the establishment of a cut score for each grade level.

Background

As part of the state assessment and accountability system, Nebraska administers Writing Assessments across the state at selected grade levels (4, 8, and 11). These assessments are used to distinguish between students who have met the Statewide Writing standards and those that have not met the Statewide Writing standards and may need additional instruction in writing. Because the Writing Assessments are used to classify students in terms of their level of performance in writing, the Department of Education has recognized the importance of using psychometrically accepted methods for setting these cut scores to operationally define performance standards.

The writing assessments give students an opportunity to provide a writing sample in response to a narrative (4th grade), descriptive (8th grade), or persuasive (11th grade) prompt. The student writes to the prompt that is provided in a given year. The prompts are scored holistically across six traits on a 10-point scale. Two trained scorers score each paper and the student's score on the paper is the sum of the two scorers' scores. If the two scorers disagree by more than one score point, a third scorer scores the response and an average of the two closest scores is computed.

The purpose of this study was to provide a recommendation for a range of defensible cut scores to the Nebraska Department of Education (NDE) for the *Statewide Writing Assessment* in grades 4, 8, and 11. This report focuses on the results of the standard-setting studies for these three grade levels. The report provides an overview of the methods and procedures for the study. It includes a recommendation for a range of cut scores within which NDE may identify a defensible cut score that will help decide which students in the state have met the writing content standards.

Methods and procedures

Overview of Procedures

Two methods for estimating a cut score were used, each of which relying on different assumptions. The use of these independent methods is intended to provide a more defensible range of possible cut scores, which NDE may use to determine the final cut score. These methods are a) an analytical judgment method (Hambleton & Plake, 2000) and b) an initial estimates method (Hofstee, 1983). These methods are described briefly below.

Each of the methods was conducted on April 29, 2008 in a workshop facilitated by staff from Alpine Testing Solutions. The workshop began with an orientation and training activity that included an extended discussion of the test specifications. The training also included a description and discussion of the following student performance levels that were developed by NDE and provided to us for use in the workshop:

1. Beginning: The writing is still under development. Extensive revision and/or editing would be necessary.
2. Proficient: The writing has more strengths than weaknesses. Some revision and/or editing would be necessary.
3. Advanced: The writing has many strengths. Only minor revision and/or editing would be necessary.

Analytical Judgment Method

One standard setting method used in the standard setting studies was a modification of a method described by Hambleton and Plake (2000). This method required panelists to read a set of 50 papers (described below) and sort the papers into the three broad performance classifications defined above (Beginning, Proficient, or Advanced). After the initial sorting was completed, panelists identified three papers from the "Beginning" papers that were the closest to being in the next higher classification (Proficient). Panelists also identified three papers classified as Proficient that were closest to being Beginning. That is, panelists identified the three best papers in the Beginning classification category and the three worst papers in the Proficient category. Panelists did not know the scores on the papers; instead each paper had an identifying code corresponding to a specific score. The cut score for a panelist was that panelist's mean of the six specific papers that were closest to next higher or lower category. The overall cut score was the average of the individual teacher cut scores.

The 50 papers were purposely selected from the total set of papers to reflect the following criteria:

1. All score points were represented by at least 2-3 papers with more papers with scores between 2- and 3+ being included.
2. Selected papers were scored correctly and accurately. The basis for scoring was not to be an issue.
3. Selected papers were written legibly and darkly enough that they could be photocopied.

Initial Estimates method

This method is a variation of Hofstee (1983) and entailed asking panelists to estimate the percentage of tested students in their classes this year who would be classified as Beginning. This was done after all training activities and before participants completed the Analytical Judgment Method. Special forms that also included demographic information to document the level of experience of the panelists were used for this method.

Specific Procedures

Analytical Judgment Method

The standard setting workshop took place in Omaha, NE at Educational Service Unit #3 on April 29, 2008. A total of 42 teachers and administrators participated with 15 at 4th grade, 14 at 8th grade, and 13 at 11th grade. All panelists were currently teaching, had recently taught English at their respective grade level, or held positions in their districts related to reading (e.g., Literacy Coach, English Supervisor, Instruction/Literacy Facilitator) and had been exposed to the six-trait writing method used to score the Writing Assessment. Some of the panelists had also participated in the scoring process and/or participated in their respective grade level's writing assessment standard setting previously.

Following introductory comments an orientation and training session was conducted. This session articulated the purpose of the standard setting workshop and detailed the steps to be taken to complete the standard setting process. Training included a discussion of the performance categories (defined above) and a discussion of each of the six traits. Marilou Jasnoch described the six writing traits to the participants and provided an overview of the training and operational scoring activities for the statewide writing assessment. After the large group orientation, the panelists were subdivided into their grade level teams for further training.

In these grade level teams, there was additional discussion of the student who was Barely Proficient as the target student. Using performance level descriptors derived from the previous standard setting workshops for the statewide writing assessment, the panelists discussed the skills and performance characteristics of the target student in each of the six traits and holistically. They added to and modified the performance descriptions to better clarify their conception of the Barely Proficient student. These revised descriptions for each grade level are included as Appendix A and could serve as starting points for discussion on future studies. Panelists were advised that they would be reading a large number (i.e. 50) of papers and would be making holistic classifications for these papers. These holistic classifications would result in three stacks of papers, those that represented work that was a) Beginning, b) Proficient, and c) Advanced.

Prior to the operational ratings, panelists were given a set of ten papers to practice the process. All panelists received the same papers to rate. These papers were selected such that there were papers that spanned the score range. Panelists made two sorting decisions using these ten papers. First the papers were classified as being Beginning,

Proficient or Advanced. After that sorting decision was made, panelists identified the paper from the Beginning papers that was closest to being in the Proficient category. They also selected the paper in the Proficient category that was closest to being in the Beginning category. After these selections were made, there was a show of hands regarding how each paper was classified. This was followed by a discussion of why panelists made their classification decision.

The training was followed by the professional judgment method and then the operational analytical judgment method. The panelists in each grade level team were provided with copies of 50 papers selected as described above. Panelists made the initial sort into the three broad categories and then selected the three best of the papers classified as being Beginning and the three papers they felt were the worst of those in the Proficient category. Papers were collected and data entered.

Initial Estimates method

After the practice analytical judgment method was completed, the Initial Estimates method was undertaken. This method entailed having the panelists estimate the percent of students in their classes this year who would be classified as being Beginning.

Results

Analytic Judgment Method

The minimum passing scores are based on the judgments of panelists who made holistic ratings on the 50 papers. Each teacher's individual cut score was computed. This involved computing both a mean of the six papers that were just above and just below the performance of the student who was "Barely" Proficient (the target student).

Grade 4

For this grade the recommended cut score using the mean was 4.10. The closest score point to this mean value would be 4.00. The panelists' recommended cut score (4.00), and a range of cut scores plus and minus 1 score point are shown in Table 1. The approximate percent of 4th grade Nebraska students who would be below the cut point is also shown in the Impact column.

For the initial estimates method, panelists' estimated percent of students who will be classified as being Beginning ranged from a low of 0% to a high of 35%, with a mean of 18.0% and a median of 20.0%. The closest score point associated with these impact values is 4.66 (impact = 18.75%).

Table 1. Analytic Judgment-based cut score and impact and cut scores and impacts within a one score point range for 4th grade.

Range	Cut score	Impact (% below)
1 Score Below	3.67	7.19
Average	**4.00**	**9.94**
1 Score Above	4.33	14.53

Grade 8

For this grade the recommended cut score using the mean was 3.8. The closest score point to this mean value would be 4.00. This cut score (4.00), and a range of cut scores plus and minus 1 score point are shown in Table 2. The approximate percent of 8th grade Nebraska students who would be below the cut point is also shown in the Impact column.

For the initial estimates method, panelists' estimated percent of students who will be classified as being Beginning ranged from a low of 0% to a high of 15%, with a mean and median of 7.00%. The closest score points associated with these impact values is 4.33.

Table 2. Analytic Judgment-based cut score and impact and cut scores and impacts within a one score point range for 8th grade.

Range	Cut score	Impact (% below)
1 Score Below	3.67	3.99
Average	**4.00**	**5.37**
1 Score Above	4.33	7.94

Grade 11

For this grade the recommended cut score using the mean was 4.10. The closest value using the mean would be 4.00. The cut score (4.00) and a range of cut scores plus and minus 1 score point is shown in Table 3. The approximate percent of 11th grade Nebraska students who would be below the cut point is also shown in the Impact column.

For the initial estimates method, panelists' estimated percent of students who will be classified as being Beginning ranged from a low of 0% to a high of 25%, with a mean of 11% and a median of 10%. The closest score point associated with the impact values was from 4.66.

Table 3. Analytic Judgment-based cut score and impact and cut scores and impacts within a one score point range for 11th grade.

Range	Cut score	Impact (% below)
1 Score Below	3.67	3.51
Average	**4.00**	**4.52**
1 Score Above	4.33	6.27

Evaluation Data

At the conclusion of the workshop, panelists completed an evaluation form consisting of four parts. Part 1 focused on the orientation and training; Part 2 addressed the panelists' levels of comfort and confidence in their Initial Estimates ratings; Part 3 was parallel to Part 2, but focused on the confidence and comfort levels for the Analytical Judgments. Part 4 consisted of closed and open-ended items asking about the overall success of the workshop and about recommended changes that might be made to improve

future workshops. Evaluation comments are shown in Appendix B. Results were similar across grade levels and are reported in aggregate across grade levels.

Part 1: Training

On a scale ranging from 1 - 6, where 1 = Very Unsuccessful and 6 = Very Successful, all mean ratings fall between 4.8 and 5.5. (Orientation mean =5.5, Training on Analytical Judgments Method mean = 5.2, Description of target students mean = 4.8, Practice with Analytical Judgments Method mean = 5.1, and Overall Training mean = 5.1).

Panelists also rated the adequacy of the time provided for training and orientation. On a six-point scale, where 1 = Totally Inadequate and 6 = Totally Adequate, all mean rating exceeded 5.1. (Orientation mean = 5.6, Training on Analytical Judgments Method mean = 5.4, Description of target student mean = 5.1, Practice with Analytical Judgments Method mean = 5.3, and Overall Training mean = 5.3).

When asked to rate the amount of time allocated to training, the mean rating was 2.0 where a value of 2 was "The right amount of time was allocated to training." One panelist that too much time was allocated to training and one felt that too little time was allocated to training.

Part 2: Initial Estimates Method

The mean panelists' confidence and comfort in making estimate using the Initial Estimates method were 3.4 and 3.5, respectively on a four-point scale (1 = Not Confident/Comfortable and 4 = Confident/Comfortable).

The mean rating for the allocation of time for making the initial estimates was 3.3 on a four point scale (1 = More time needed to be allotted to complete this judgment and 4 = More than enough time was allotted to complete this judgment).

Part 3: Analytical Judgments Method

The mean panelists' confidence in classifying papers into three categories was 3.7 on a four-point scale (1 = Not Confident and 4 = Confident). The mean Comfort rating on the same 4-point scale (1= Not Comfortable and 4= Comfortable) for this process was also 3.7.

The final item in Part 3 asked about the adequacy of time allocated for making the analytical judgments. On the four-point scale (1 = More time needed and 4 = More than enough time was allotted), the mean rating was 3.7.

Part 4: Overall Workshop Ratings

The first item in Part 4 asked about the panelists' confidence in the passing standard that would result from this process. The mean confidence was 3.6 on a four-point scale (1 = Not at all Confident and 4 = Confident). Thus, overall panelists were "Confident" about the appropriateness of the passing standard. None of the panelists had a confidence rating less than 3.

Two questions asked panelists to rate the success and organization of the workshop (1 = Totally Unsuccessful and 4 = Totally Successful). The mean ratings on these items were both 3.5.

Panelists were also given an opportunity to provide comments they felt would be helpful in planning future standard setting studies. These comments are attached in Appendix B.

Conclusions and Recommendations

The panelists' recommendations for each grade level are based on considerations of both methods described in the body of this report. For each grade, the cut score based on the Initial Estimates method was higher than the Analytical Judgment method. Although the results of both methods are reported here, we believe there is greater stability in the recommendations produced by the Analytical Judgment method because the number of teachers who provided Initial Estimates may not sufficiently generalize to Nebraska's classrooms as a whole.

For 4th grade, the Analytical Judgment method produced a recommended cut score of 4.00; whereas the Initial Estimates method yielded a recommended value of 4.66. If a cut score of 4.00 is adopted, approximately 9.9% of Nebraska's 4th grade students would be identified as Beginning. If the higher value was used, 18.8 percent of students would be classified as Beginning.

At 8th grade, the Analytical Judgment method also produced a recommended cut score of 4.00. The Initial Estimates method yielded a cut score 4.33. If a cut score of 4.00 is adopted, approximately 5.4% of Nebraska's 8th grade students would be identified as Beginning. Using the higher value, approximately 7.9% of students would be classified as Beginning.

Finally, for 11th grade, the Analytical Judgment method also produced a recommended cut score of 4.00. The Initial Estimates method yielded a recommended cut score of 4.66. If a cut score of 4.00 is adopted, approximately 4.5% of Nebraska's 11th grade students would be identified as Beginning. Using the higher value, approximately 8.1% of students would be classified as Beginning.

References

Hofstee, W. K. B. (1983). The case for compromise in educational selection and grading. In S. B. Anderson and J. S. Helmick (Eds.), On educational testing (pp. 109-127). Washington, DC: Jossey-Bass.

Plake, B. S., & Hambleton, R. K. (2000). A standard-setting method designed for complex performance assessments: Categorical assignments of student work. Educational Assessment, 6(3), 197-215.

Appendix A

Performance Level Descriptions for Grades 4, 8, and 11

High School (Grade 11): Defining Proficiency for the 6 Traits

Advanced

Stronger	Weaker
Word Choice	
Organization	
Conventions	
Ideas (original)	
Appropriate Voice	
Sentence Fluency	

Proficient

Stronger	Weaker
Ideas (safe, may lack originality)	Voice
Organization	Sentence Fluency
"Word Choice" and "Conventions" - somewhere between strong and weak	

Beginning

Stronger	Weaker
May attempt "Word Choice"	Organization
Voice	Sentence Fluency
	Conventions
	Ideas – sketchy

Proficient 11th grader

Strengths	Weaknesses
Ideas	
Has details	Not fully developed
Commitment to the writer's position	
Clear	
Organization	
Functional paragraphing structure	
Sequencing logical	Sequencing not always complete
Evidence of Beginning, Middle, and End	Transitions not always there
Voice	
Some conviction	Sometimes forced or stilted language
May observe some individuality and naturalness Appropriate tone for the audience	
Word Choice	
Clear and somewhat persuasive – appropriate for 11th grade	Some trite, non-specific language
Most words used correctly – no undue confusion Specific language	Language not always used appropriately in context
Sentence Fluency	
Some variety in structure and length	Phrasing may be more mechanical (related to flow) with little variety
Some transitions	Run-ons, fragments
Mostly flowing	
Conventions	
Errors don't detract from readability (some editing)	Few correct uses of stylistic punctuations.
Basic punctuation, usage	Errors in spelling and grammatical usage may distract, but not confuse

Grade 8: Defining Proficiency for the 6 Traits

Advanced

Stronger	Weaker
Word Choice	
Organization	
Conventions	
Ideas	
Voice	
Sentence Fluency	

Proficient

Stronger	Weaker
Word Choice	Voice
Organization	Sentence Fluency
Conventions	
"Ideas" - somewhere between strong and weak	

Beginning

Stronger	Weaker
May attempt "Word Choice"	Organization
Voice	Sentence Fluency
	Conventions
	Ideas – sketchy

Proficient 8th grader

Strengths	Weaknesses
Ideas	
Relevant to topic with some details	Not fully developed
Clear	Not always apparent
Organization	
Sequencing logical	Not always complete
Evidence of Introduction, Body & Conclusion	Transition not always there
Transitions when attempted are predictable	
Voice	
Some sense of personality Some audience consideration	Sometimes forced or mechanical
Word Choice	
Word used correctly	Word choice may not be creative Some trite, non-specific language
Some sensory details are apparent Some specific words	Not fully developed or used appropriately in context
Sentence Fluency	
Some variety in structure and length	Phrasing may be more mechanical (related to flow)
Mostly flowing	Needs more development
Conventions	
Errors may occasionally detract from readability (significant editing)	Do not expect stylistic punctuation (e.g., hyphens, quotation marks)
Basic capitalization and end punctuation.	
Reader can still understand the message	
Some paragraphs	

Grade 4: Defining Proficiency for the 6 Traits

Advanced

Stronger	Weaker
Word Choice	
Organization	
Conventions	
Ideas	
Voice	
Sentence Fluency	

Proficient

Stronger	Weaker
Word Choice	Voice
Organization	Sentence Fluency
Conventions	

"Ideas" - somewhere between strong and weak

Beginning

Stronger	Weaker
May attempt "Word Choice"	Organization
Voice	Sentence Fluency
	Conventions
	Ideas – sketchy

Proficient 4th grader

Strengths	Weaknesses
Ideas	
Has some details	Not fully developed
May attempt creativity	Not always apparent
Clear	Multiple off topic details
Organization	
Sequencing logical	Not always complete
Evidence of Beginning, Middle, and End	Transition not always there, not always logical
Hook & Conclusion are attempted	
Voice	
Some personality, evokes some feeling	Sometimes forced or mechanical
Word Choice	
Clear and somewhat descriptive – appropriate for 4th grade	Some trite, non-specific language
Sensory details may be attempted	Not fully developed or used appropriately in context Unnatural, exaggerated choice of words
Sentence Fluency	
Some variety in structure and length	Phrasing may be more mechanical (related to flow)
Some transitions even if basic	Needs more development
Mostly flowing	
Conventions	
Errors don't detract from readability (some editing)	Few attempts to use stylistic punctuations.
Basic punctuation, usage (Uses ending punctuations -- . ? !)	
Reader can still understand the message. Attempt at paraphrasing.	No attempts at paraphrasing

Appendix B. Comments from Standard Setting Workshop Evaluation

- Great! Efficient!
- At the top of the bright pink sheet - write "proficient level" so we know that is the target for that side.
- Very Well Organized! Thanks!
- Our facilitator was nice and pleasant to work with, but did not seem confident with leading the process (this could be due to someone in our group trying to run things).
- I feel our presenter could have been trained a bit more.
- I found this very helpful as a 4th grade teacher.
- I would like to see the breakdown a little more (how does a score get figured between a "high beginning" and a "low proficient").
- If our trainer was more confident, I would feel more secure.
- Very interesting, and it felt good to be involved on the other side of things! I would do it again.
- Our individual group leader was not confident. Luckily, because of our group's knowledge, we were very close in the decisions in the practice rounds.
- I just need more practice! I'll be back!
- The trainers did a fantastic job.
- I thought the day went smoothly. Thanks for letting me participate. And great meals and facilities!
- Thank you for treating us so well! It is always a pleasure to be here.
- The clarification of the Proficient Student on the Pink Sheet helped!
- This was my first workshop of this nature and I would give it a very positive rating! The training was useful and stimulating. I feel like I learned some good things and hopefully helped out.
- Thank you for this experience!
- Thanks for the opportunity to participate in the process!
- The workshop provided excellent professional development in terms of the process. I have a much better understanding of why the cut score for proficiency is lower than it seems it should be. In other words, the definition of proficiency is designed as lower than it is on the scoring guide (6) or each rater assigning an average of 3.