AN ANALYSIS OF THE
MICHIGAN EDUCATIONAL ASSESSMENT PROGRAM RUBRIC
AND THE
NATIONAL WRITING PROJECT ANALYTICAL WRITING CONTINUUM
by
Jolene M. Hetherington

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF MASTER OF ARTS IN ELEMENTARY EDUCATION
AT NORTHERN MICHIGAN UNIVERSITY

April 15, 2010

APPROVED BY:

DATE:
# TABLE OF CONTENTS

List of Tables .................................................................................................................. 4

Abstract .......................................................................................................................... 5

Chapter One: Introduction............................................................................................... 6

  Scoring Rubric Development and Use ........................................................................... 6

  Statement of Problem ............................................................................................. 7

  Research Questions ............................................................................................. 8

  Definition of Terms ............................................................................................. 8

Chapter Two: Literature Review ................................................................................... 11

  Assessing Writing .............................................................................................. 11

  Scoring and Interpreting Process on State Tests and in the Classroom............... 13

  Teaching the Rubric........................................................................................... 14

  Quality of Writing ............................................................................................. 16

Chapter Three: Methodology ......................................................................................... 20

  Sampling Plan ................................................................................................... 20

  Participants ........................................................................................................ 20

  MEAP Writing Assessment ............................................................................... 20

  Inquiry Project and Training .............................................................................. 21

Chapter Four: Results .................................................................................................... 23

Chapter Five: Conclusions ............................................................................................ 28

  Discussion ......................................................................................................... 28

  Recommendations ............................................................................................. 29

  Further Study ..................................................................................................... 30
Conclusion ........................................................................................................ 32
Acknowledgements ........................................................................................... 32
References .................................................................................................................... 33

Appendix A: Michigan Educational Assessment Program (MEAP) Writing from
Knowledge and Experience Grades 3-8 Holistic Scorepoint Descriptions.................... 38

Appendix B: Comment Codes: ELA Writing from Knowledge and Experience............ 39
LIST OF TABLES

Table 1: Characteristics of Students Who Wrote Papers (n=225) ................................... 20
Table 2: Frequency of Distribution of Scores for the MEAP and AWC ......................... 23
Table 3: Distribution of MEAP Comment Codes ........................................................... 24
Table 4: Interrater Agreement for MEAP and AWC by Grade and Score ..................... 25
Table 5: Dependent Sample t-test of Mean Scores for MEAP and AWC ....................... 25
Table 6: Correlation Study of MEAP and AWC ............................................................ 26
Table 7: Point Biserial Correlation between Comment Codes and MEAP or AWC ....... 27
Abstract

Writing assessments are a required component of many state assessments. Rubrics provide a valid model to assess writing with accuracy. The MEAP rubric and the NWP’s Holistic rubric and AWC were evaluated using student writing produced during the 2008 MEAP Writing Assessment. The results suggest that the holistic rubrics of the MEAP and AWC holistic rubrics are correlated but the MEAP comment codes do not correlate with the MEAP rubric nor the AWC. The NWP AWC provides more information to help teachers help students write more effectively than the MEAP rubric.
Chapter One: Introduction

Good writing matters because in the twenty-first century the written word remains the principal way of communication (Hamp-Lyons, 2002). The difficult part in writing assessment is determining what is good writing, what qualities make it good, and who actually makes the determination of the quality of writing.

Scoring Rubric Development and Use

Scoring rubrics are used for a variety of reasons and “are typically employed when a judgment of quality is required” (Moskal, 2000a). The purpose of the rubric depends on the purpose of assessment. A common use of scoring rubrics is in the assessment of writing. In developing a rubric, criteria need to be established to make the subjectivity involved more objective. A rubric is a document that articulates the expectations for an assignment by listing criteria, or what counts, and describing levels of quality from excellent to poor. Rubrics evaluate whether criteria have been reached and provide feedback to students and teachers on their efforts and how to improve the writing. Rubrics allowed quick and efficient scoring for teachers and help justify grades to parents and others (Andrade, 2000).

The NWP developed the Analytic Writing Continuum (AWC) (National Writing Project, 2006) by adapting the 6 + 1 Traits of Writing model (Culham, 2003). Although the 6 + 1 Traits of Writing is sufficiently comprehensive, it required modifications to make it more appropriate for research purposes. Local writing project sites can use the AWC to assess student writing as well as teacher consultants who may use in their classroom only if they have been trained in its use. The NWP rubrics are not to be distributed by anyone outside formal site sponsored professional development activities (National Writing Project, 2008, June).
Michigan Educational Assessment Program (MEAP) assesses students’ writing grades 3-8 based on Michigan Curriculum. The results of the MEAP writing assessment are presented using a rubric score from 1 to 6 and are supplemented by comment codes that provide feedback from MEAP scorers.

Both the MEAP and the NWP use holistic scoring which is a method that trained readers evaluate a piece of writing for its overall quality. The NWP holistic scoring requires readers to evaluate the work as a whole, while considering the major elements of content, structure, stance, sentence fluency, diction, and conventions. The MEAP holistic rubric considers similar elements. In this type of scoring, readers are trained not to become overly concerned with any one aspect of writing but to look at a response as a whole. Holistic scoring is not a summation or averaging of the analytic scores, but a reflection of the overall effect of writing (National Writing Project, 2008).

Analytic scoring is designed to rate papers on separate criteria. Where holistic scoring assigns a single score that represents a weighting of all the different attributes addressed by the scoring guide, an analytic score might assign a high score to one attribute, such as “structure,” and a lower score to another, such as “conventions.” Analytic scoring may be slower than holistic, but provides more feedback. The analytic elements that make up the AWC are the scoring categories (National Writing Project, 2008).

Statement of Problem

When the MEAP results are sent back to local schools, teachers desegregate the data to improve their teaching and their students’ learning. Teachers have expressed frustration with the type of feedback given on the test results including the comment codes. The 6-point holistic rubric does not break down the difficulty with the different
attributes of writing. Nor do the comment codes give teachers feedback since some of the attributes are combined into one comment code. (See Appendix B, Comment Code 2.)

Teachers trained in the AWC discovered that using the analytical rubric gave them more feedback when working with their students. They could focus on the attribute(s) of writing their students struggled with and develop lessons.

Research Questions

Considering the two rubrics, the following questions were developed:

1. What are the correlations between the MEAP holistic rubric and the AWC holistic rubric?

2. Do the comment codes correlate with the AWC analytical rubric attributes as well as the holistic scores?

These questions will be answered using constructivist theoretical approach.

Definitions of Terms

AWC attributes.

Content (including quality and clarity of ideas and meaning)—The content category describes how effectively the writing establishes and maintains a focus; selects and integrates ideas related to content (i.e., information, events, emotions, opinions, and perspectives); and includes evidence, details, reasons, anecdotes, examples, descriptions, and characteristics to support, develop, and/or illustrate ideas.

Conventions—The conventions category describes how effectively the writing demonstrates age-appropriate comparison of usage, punctuation, spelling, capitalization, and paragraphing.
Diction (Language)—The diction category describes the precision and appropriateness of the words and expressions for the writing task and how effectively they create imagery, provide mental pictures, or convey feelings and ideas.

Sentence Fluency—The sentence fluency category describes how effectively the sentences are crafted to serve the intent of the writing in terms of rhetorical purpose, rhythm, and flow.

Stance—The stance category describes how effectively the writing communicates a perspective through an appropriate level of formality, elements of style, and tone appropriate for the audience and purpose.

Structure—The structure category describes how effectively the writing establishes logical arrangement, coherence, and unity within the elements of the work and throughout the work as a whole. (National Writing Project, 2008, June; Swain, Graves, & Morse, 2007)

Other terms.

Adjacent Scores – When two scores assigned to a writing sample are within one score point of each other; for example, a paper given a “4” by one scorer and a “3” by another. (Michigan State Department of Education. Michigan Educational Assessment Program, 1997).

Analytical Scoring – The method by which trained readers evaluate a piece of writing by rating on separate criteria. (National Writing Project, 2008, June)

Anchor Papers (also called benchmark papers) – Student-produced writings on a specific topic that exemplify the score standards for each level on the scoring guide. (Michigan State Department of Education. Michigan Educational Assessment
Holistic Scoring – The method by which trained readers evaluate a piece of writing for its overall quality. (National Writing Project, 2008, June)

Non-adjacent Scores – When two scores assigned to a writing sample are not exact or within one score point of each other; for example, a paper given a score of “4” by one scorer and a “2” by another. (Michigan State Department of Education. Michigan Educational Assessment Program, 1997).
Chapter Two: Literature Review

Assessing Writing

Rubrics have essentially two uses in a classroom: “1) to gather information on students in order to plan instruction, track student progress toward important learning targets, and report progress to others; and 2) to help students become increasingly proficient on the very performances and products also being assessed.” (Arter, 2000). Rubrics also allow consistency in scoring and improved instruction for teachers and allow students to know what they will be assessed on (Arter, 2001; Trumbull & Farr, 2000).

Rubrics tend to improve interrater reliability, meaning different scorers will give similar scores, because rubrics limit the scope of the variability of scores (Mabry, 1999). Writing assessment is an exercise of balancing validity and reliability. Validity refers to measuring what is intended to be measured whereas reliability is measuring it consistently (Yancey, 1999). Content, construct, and criterion are examined to support validity of writing assessments (Moskal, 2000b). The reliability of assessments can be strengthened by using tighter restrictions on the writing assessment and the use of rubrics can improve reliability in most cases (Jonsson & Svingby, 2007).

Mabry (1999) stated that standardizing scoring with a writing rubric presents a validity problem because rubrics incorrectly imply that good writing is the sum of the criteria on the rubric and that writing that does not meet the criteria is not good. However, results indicated that the New Zealand government and teachers could have confidence in the validity of the country’s writing assessment (Brown, Glasswell, & Harland, 2004).

Brown, Glasswell, and Harland (2004) conducted two studies on the validity of the New Zealand scoring system used in classrooms, In Study 1, seventeen experience classroom teachers
who had no experience in scoring large scale assessments met and scored writing samples. They scored six different writing prompts with an average of 589 samples each. The teachers’ training included a 1.5- hour grammar instruction lecture, scoring rubric overview, and specific training on the progress indicators for describe/report writing. In the analysis of the scoring, the interrater reliability was assessed on three components: consensus, consistency, and measurement approaches. Consensus was measured when the score agreement occurred. Score agreement was understood to occur when the scores were equal or fell into the range of +1 or -1 of the expert scorer’s score. The close agreement rates ranged from 66% to 92%. Consistency was measured by analyzing the SPSS coefficient alpha reliability routine. These alpha coefficient estimates ranges from .53 to .87. The Brennan and Kane dependability index assessed the measurement approaches. Values ranged from .67 to .95 which showed that the scores approached the threshold for dependable values for the most part.

In Study 2, over 60 students from three high schools completed an argumentative writing training program which included explicit instruction and modeling, scaffolding, and metacognition monitoring. The four classrooms were taught how to write persuasively for an average of 27.25 hours and concluded with a two-hour assessment. The consistency estimate of reliability was very at .95 for the assessment. The study also had three writing specialist rescore 42 writing samples with no training. Consistency coefficients of agreement between the two types of assessors averaged r=.70 (SD=.05) for deep scores and r = .60 (SD = .07) for surface scores. Another important finding is this study was the huge gain in quality of persuasive writing after the training program. The gain was equivalent to more than three years’ average progress. The two studies showed adjacent of agreement of 70-90% with .70-.80 consistency and measurement correlations.
In assessing a type of writing instrument, it is vital to make sure that the assessment is valid and reliable. Both the MEAP and the NWP rubrics have been shown to be both.

**Scoring and interpreting process on state tests and in the classroom**

When scoring with rubrics, two independent scorers should obtain consistent scores on the same writing sample and scorers should obtain consistent scores over time (Moskal, 2003). To achieve this, calibration of scorers is critical. To calibrate a scorer, the rater scores the anchor papers based on the rubric(s) provided. Then, as a group, the scorers discuss the results and then the anchor paper score is revealed. The scorers must agree that the anchor paper score is correct in order for calibration to work. As writing samples are scored, the scorers may look back at the anchor papers for assistance.

The rating of student writing depends upon the anchor papers used in scoring. Narrative third grade writing samples were scored higher when the anchor papers were of the same grade whereas they were scored lower when scored against anchor papers from grades 5 and 8 (Popp, Osborn, Thompson, & Behrens, 2003). Grades 3, 5, and 8 responded to writing prompts with each grade having its own type of prompt: Grade 3 – narrative prompt, Grade 5 – literary response prompt, and Grade 8 – persuasive essay prompt. Grade 3 contained 317 samples. Approximately 15% of students in Grade 5 (180 students) and Grade 8 (172 students), also responded to a narrative prompt. Twelve different raters scored the across-grades samples for the narrative prompts and six or seven raters scored the within-grade writing samples which each sample being read and by two different raters. Raters used a six-point rubric similar to 6+1 Traits of Writing (Culham, 2003.) and used anchor papers to guide the scoring. The results showed that papers rated with with-in grade level anchor papers the mean of the summed scored points were 20.7 (SD=3.76). The papers scored with the across-grade level anchor papers scored
17.0 (SD =4.32). Popp et al. (2003) also found that the rating of student writing increases with grade level. Students in higher grades performed better than students in lower grades when given the same writing assignment.

The NWP National Scoring Conference has scored 32,672 papers. Of those, 16% to 23% of the papers were doubled scored. The reliabilities of the seven scores (a holistic and six analytic) ranged from 87% to 92% with a mean of 90% (LeMahieu, Swain, Murphy, & Sperling, 2009). Although Burns (1998) conducted a study to interpret the reliability and validity of the MEAP, the only writing assessment result was the High School Proficiency Test at .674 and .654 for two different forms of the test. On the 2008 MEAP writing assessment, both a scorer and a team leader read 20% of the writing books to check for accuracy. An overall agreement rate of 78.3% was maintained between the two groups (Wang, Wang, & Viger, 2008).

**Teaching the rubric**

As rubrics become more common in the classroom, Andrade, Du, and Wang (2008) looked at the impact of using a rubric to teach as well as to evaluate. Their quasi-experimental study (Patten, 2007, p 5) investigated the effect of using good writing models, developing a list of criteria for quality writing, and having students self-assess their first drafts using a rubric. Seven public elementary classrooms took part in the study with three third grades (N=46) and four fourth grades (N=70) from Northeastern United States for a total of 116 participants. Five of the seven classrooms took part during their ELA time while the other two participated during their social studies time. Participants were primarily Caucasian (N=110) with 52 males and 64 females. Only three of the participants were identified as having special needs.

The treatment group and the control group received a writing assignment. Six classes wrote persuasive essays with four writing on the topic of year-round schooling and the other two
[a treatment group and a control group] focused on the impact of European settlers on Native Americans. A third grade class, a treatment group, wrote stories about their families. The writing topics had to relate to the curriculum, as that was a condition of participation by the teachers. The treatment group differed from the control group in three different ways: The treatment group read and discuss a model story or essay, discussed the written rubric, and self-assessed their first drafts using that rubric. Both groups were introduced to the writing assignment, generated a list of qualities of an effective story or essay, did prewriting, wrote a first draft, and self-assessed the first draft with the control group just reviewing their first drafts and noting possible changes for final draft.

Andrade, Du, and Wang collect several measurements of data. The first author recorded time spent on writing. Students’ previous English/Language Arts achievement was based on recent ELA grades. Questionnaires were use to measure the recent prior use of a rubric. Six researchers scored the essays blind to the treatment condition using rubrics adapted from the rubrics used in the treatment classrooms. The six researchers tested the rubrics by scoring a series of essays together and revised the rubrics repeatedly. A scoring procedure was developed to manage the scoring done by each pair of scorers and to track inter-rater reliability.

A GLM two-way ANOVA indicated that the treatment group’s writing scores (M = 28.5, SD = 4.9) were higher than the control group’s scores (M = 24.3, SD = 4.7). Females were found to higher essay scores (M = 27.7, SD = 5.1) than males (M = 25.8, SD = 5.2) but that statistically the scores do not reach significance. ANOVA was also used to analyze the results excluding the story scores. Andrade, Du, and Wang also examined the relationships between the treatment and the seven criteria of ideas and content, organization, paragraph formatting, voice and tone, word choice, sentence fluency, and conventions. Controlling for previous achievement, the treatment
has a statistically significant relationship with the criteria scores. For the other data collected, the mean of each was calculated.

Andrade, Du, and Wang concluded that having students use model papers to generate criteria for good writing and use a rubric for self-assessment relates positively to the quality of writing produced. Females tended to produce higher scoring writing than males. Time spent on writing and prior rubric use did not correlate with the scores received. The authors also discovered that the treatment had a significant influence on the seven criteria with the exception of sentences and conventions.

Quality of writing

As state assessment offices reduce the amount of time allowed for writing assessments, Crawford and Smolkowski (2008) looked at the quality of on-demand writing. Their correlational research study (Patten, 2007, p 9) examined the difference between scores on the first draft and the final draft of a state writing assessment. Based on students’ previous year’s proficiency scores on the state assessment, a stratified random sampling plan was used to collect a total of 172 fifth-grade writing samples and 176 eighth-grade writing samples. The samples were collected from eight elementary schools and three middle schools from a relatively large urban school district (N=19,894) located in the western United States. Also included in the samples were twenty additional tests completed by special education students. Five of the fifth grade samples were eliminated due to: 1) changes in three drafts not being able to be tracked, 2) a draft using a scribe to accommodate a special education student, and 3) a first draft seemed be written by a different author than the final draft.

The fifth grade participants were 97 males and 70 females; 127 general education students and 40 special education students; 137 Caucasian, 5 American Indian, 10 Hispanic, 4
African-American, and 11 Asian/Pacific Islander. The eighth grade participants were 124 males, 51 females, and one unknown gender; 130 general education students and 46 special education students; and 143 Caucasian, 15 Hispanic, 10 African-American, and 8 Asian/Pacific Islander.

Students were given two sessions to complete the extended response section of the state assessment. In the first 50-minutes session, students planned and wrote a first draft. In the second 50-minutes session, the students were given time to revise, edit, and rewrite their composition in a ‘clean’ booklet. The first and final drafts were copied and transcribed and then double-checked for accuracy.

A team of six current or former elementary or middle school teachers attended two days of training for four hours each day. At the training, the team reviewed the trait rubric, read ten identical samples one at a time and discussed the scoring of each trait extensively, made minor revisions to rubric to clarify scores, took home ten additional samples to score at home individually and collected individual scores which were calculated inter-rater agreement. From the results of this process, a third day of training was scheduled to provide the scoring team additional feedback. After this additional training, four scorers were either assigned a first draft or a final draft of each paper so that a scorer did not score all the first drafts or the final drafts or score both the first and final drafts of the same student. Scorers were blind to the purpose of the study and unaware if the draft in front of them was a first or a final draft. The scorers using the rubric with the three categories of style/fluency, content/organization, and language usage read each draft and gave each sample a total score. The last two of the six scorers were considered readers and were assigned papers when the scores from the two scorers did not agree.

Two analysis approaches were employed to disaggregate data by grade and educational classification. The analyses tested for changes in scores between the first and final drafts by
subtracting the first draft score from the final draft. The change scores were then compared to zero with t-tests (Patten, 2007, p. 125-6). Two-tailed p-values were used to allow for inferences in the student scores decreased. The t-tests could detect effects of .25 with the power of .80 and .32 with the power of .95 with a value for alpha of .05 in fifth grade and eighth grade respectively for the general education samples. In the smaller special education samples the t-tests could detect effects of .45 with power of .80 and .58 with a power of .95. Scores were also analyzed using the Wilcoxon matched-pairs signed-rank test to investigate the differences within a dependent sample (Patten, 2007, p 11). Finally, cross-tabulations were calculated to discover the percentages of writing sample quality that increased, decreased, or remained neutral during the final draft. In both grades, no significant differences were found between the special education first and final drafts. On the fifth grade final drafts, 25% of students received lower scores, 29% of students received neutral scores, and 46% received higher scores. On the eighth grade samples, 31.6% scored lower, 28.4% scored neutral, and 40% scored higher on the final drafts.

Crawford and Smolkowski (2008) concluded that special education students have no significant differences between the first and final drafts during an on-demand state writing assessment. However, the small special education sample may make it difficult to detect those differences. The eighth grade level also found no significant differences between the two drafts. It was also very noteworthy that only 40% of eighth graders scored higher on the final draft. In contrast, the fifth graders scored significantly better on the final draft on two out of the three traits and on the total score. After collapsing the scores for the general education and special education students, 54% of the first draft received the same or higher score than the final drafts.
Assessing by rubrics is prevalent in the world of writing assessment in many states and around the world. Many different rubrics are used but the important aspect is consistency in its use and the feedback that is given after the assessment. In the following analysis, the MEAP rubric and the NWP’s AWC holistic and analytic rubric are investigated.
Chapter Three: Methodology

Sampling Plan

The writing samples from a large rural district in the northern Michigan were randomly selected for Grades 3, 4, and 5 from the MEAP Writing CDs containing every writing sample assessed by the State for that district and were assigned a uniquely identifying number.

Participants

The 225 writing samples included in this study were completed by third-, fourth-, and fifth graders. Among the participants, 106 (47%) were females and 119 (53%) were boys. Due to the population of the district, the majority of the students were white (90%) with the next significant population being Native American (8%). See Table 1 for characteristics of the students who wrote the samples.

Table 1. Characteristics of Students Who Wrote Samples (n=225).

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Attribute</th>
<th>Number of Writing Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade level</td>
<td>Third</td>
<td>75 (33%)</td>
</tr>
<tr>
<td></td>
<td>Fourth</td>
<td>75 (33%)</td>
</tr>
<tr>
<td></td>
<td>Fifth</td>
<td>75 (33%)</td>
</tr>
<tr>
<td>Gender</td>
<td>Female</td>
<td>106 (47%)</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>119 (53%)</td>
</tr>
<tr>
<td>Ethnicity/Race</td>
<td>Asian</td>
<td>2 (&lt;1%)</td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td>3 (1%)</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>1 (&lt;1%)</td>
</tr>
<tr>
<td></td>
<td>Native American</td>
<td>17 (8%)</td>
</tr>
<tr>
<td></td>
<td>White</td>
<td>202 (90%)</td>
</tr>
</tbody>
</table>

MEAP Writing Assessment

Writing assessments are a required component of many state assessments. Depending on the state, students can have one to three testing sessions where they are prompted to plan, revise, and edit. In 2008, the State of Michigan tested students in Grades 3-8 using its MEAP English Language Arts [ELA] Assessment over a period of two days. A “Writing from Knowledge and
Experience” prompt was given on the first day of testing (Michigan State Department of Education, 2008a).

**Inquiry Project and Training**

The Upper Peninsula Writing Project [UPWP] developed an inquiry project to determine the correlation, if any, between the MEAP rubric with its comment codes and the National Writing Project’s Analytic Writing Continuum [AWC]. In June 2009, twelve UPWP Teacher Consultants [TCs] gathered to score writing samples from the 2008 MEAP Writing Assessment to collect data. The TCs were trained by a representative from NWP who is an expert in the use of the holistic rubric and the six attributes - content, structure, stance, sentence fluency, diction, and conventions.

The training began with TCs signing an agreement that the NWP’s AWC would not be shared with those who have not been trained to use them. Then the trainer presented an overview of the holistic rubric and the six attributes discussing the differences between the six levels of the rubrics. TCs were given anchor papers along with the scores and commentary for the holistic rubric and attributes. The trainer led the discussion point-by-point as to why each score was given. TCs were calibrated to the scoring system using a new set of anchor papers. Scores were first determined holistically and then for each AWC attribute.

After training and calibration, TCs were split into two groups to evaluate Michigan Educational Assessment Program “Writing from Knowledge and Experience” samples for Grades 3, 4, and 5. The writing samples each had their own unique identification number. The first group scored third and fourth grade samples while the other looked at fifth grade samples. Each TC was also given a unique number for identification purposes and scored the writing samples based on the holistic rubric first and then the attributes. Sixteen percent of the writing
samples were double scored for reliability. The scores were then entered into the computer. The initial results were back scored by the NWP representative to make sure scorers were correctly calibrated. The results were then turned over to the data team who entered the results on the computer. Any samples where the scores were not identical or adjacent were adjudicated.

After the scoring conference ended, the data collected was sent to the NWP Research and Evaluation to be evaluated and analyzed. The results of this analysis are presented in this paper.
Chapter Four: Results

The raw data from the UPWP inquiry project were transmitted to the National Writing Project along with the MEAP rubric scores reported by the state and the comment codes for each writing sample. An initial set of frequency distribution was run to examine the pattern and levels of scores assign by the two systems.

In Table 2, the frequency of distribution of the MEAP and AWC holistic rubrics are presented. The percentages for each grade level are not included because the \( n \) is identical for each grade level. The AWC holistic uses the full range of scores from either 1 to 5 or 1 to 6 while the MEAP only ranges from 1 to 4. This is reflected in the slightly lower mean scores for the MEAP. The greatest discrepancy of score frequency occurs at Grade 4. MEAP scored 15 papers at a 1 where as AWC holistic scored only two papers at that level. The score of 2 was given by 31 MEAP scorers and 12 AWC holistic scorers.

<table>
<thead>
<tr>
<th>Scoring System</th>
<th>Score Frequency</th>
<th>Mean Score of Writing Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 3 (n=75)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEAP</td>
<td>12 34 25 4 0 0</td>
<td>2.28</td>
</tr>
<tr>
<td>AWC Holistic</td>
<td>11 22 31 6 5 0</td>
<td>2.63</td>
</tr>
<tr>
<td>Grade 4 (n=75)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEAP</td>
<td>15 31 19 10 0 0</td>
<td>2.32</td>
</tr>
<tr>
<td>AWC Holistic</td>
<td>2 12 28 17 12 4</td>
<td>3.49</td>
</tr>
<tr>
<td>Grade 5 (n=75)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEAP</td>
<td>7 22 35 11 0 0</td>
<td>2.67</td>
</tr>
<tr>
<td>AWC Holistic</td>
<td>14 12 28 16 5 0</td>
<td>2.81</td>
</tr>
</tbody>
</table>

The distribution of MEAP comment codes (see Table 3) determined that there were no big differences between the grades. Each writing sample could receive up to two comment codes with eighteen samples receiving no comment code at all. None of the writing samples received
comment code #8 as no writing samples scored a 6 from the MEAP scorers. Most of the writing samples received MEAP comment code #3 (76%) – Needs details and examples to adequately develop the ideas and content. Other codes, such as MEAP comment code #2 (2%), demonstrate limited control over sentence structure, vocabulary, and/or conventions, were rarely used. These comment codes are explored later in regards to their relationship with the scores (see Table 7.)

Table 3. Distribution of MEAP Comment Codes.

<table>
<thead>
<tr>
<th>MEAP Comment Code</th>
<th>Number (and percentages (^a)) of samples that received comment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
</tr>
<tr>
<td>#1 Lacks focus on one central idea.</td>
<td>25 (11%)</td>
</tr>
<tr>
<td>#2 Demonstrates limited control over sentence structure, vocabulary, and/or conventions.</td>
<td>4 (2%)</td>
</tr>
<tr>
<td>#3 Needs details and examples to adequately develop the ideas and content</td>
<td>170 (76%)</td>
</tr>
<tr>
<td>#4 Lacks coherent organization and/or connections between ideas</td>
<td>56 (25%)</td>
</tr>
<tr>
<td>#5 Needs richer development of central idea with some additional, relevant details and examples.</td>
<td>18 (8%)</td>
</tr>
<tr>
<td>#6 Needs tighter control of organization and/or the connections among ideas.</td>
<td>6 (3%)</td>
</tr>
<tr>
<td>#7 Need greater precision and maturity of language.</td>
<td>10 (4%)</td>
</tr>
<tr>
<td>#8 Earned the highest score of 6.</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>No Comment Code</td>
<td>18 (8%)</td>
</tr>
<tr>
<td>Total Comment Codes</td>
<td>289</td>
</tr>
</tbody>
</table>

\(^a\) Column entries (number and percentage) represent the papers that received each comment in turn. They do not add up to 100% as scoring procedures allowed papers to receive a second comment code.
Interrater agreement is defined as identical or immediately adjacent scores assigned by two scorers. The holistic interrater agreements shown in Table 4 for both the MEAP and AWC

<table>
<thead>
<tr>
<th>Scoring System</th>
<th>Holistic</th>
<th>Content</th>
<th>Structure</th>
<th>Stance</th>
<th>Sentence Fluency</th>
<th>Diction</th>
<th>Conventions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 3 MEAP</td>
<td>98%</td>
<td>83%</td>
<td>92%</td>
<td>92%</td>
<td>92%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>AWC</td>
<td>92%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 4 MEAP</td>
<td>97%</td>
<td>92%</td>
<td>92%</td>
<td>100%</td>
<td>92%</td>
<td>67%</td>
<td>83%</td>
</tr>
<tr>
<td>AWC</td>
<td>92%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 5 MEAP</td>
<td>94%</td>
<td>83%</td>
<td>83%</td>
<td>92%</td>
<td>100%</td>
<td>100%</td>
<td>92%</td>
</tr>
<tr>
<td>AWC</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

by grade and attribute are in the 90% range and show that both systems have adequate reliabilities. Using the above definition, the observed interrater agreement for the AWC holistic and six attributes are as follows: holistic 94%, content 86%, structure 86%, stance 94%, sentence fluency 94%, diction 89%, and conventions 92%. For the MEAP, 20% of the writing samples were double scored to permit estimates of interrater reliability whereas 16% were double scored using the AWC holistic rubric and attributes.

Table 5. Dependent Sample t-test of Mean Scores.

<table>
<thead>
<tr>
<th>Scoring System</th>
<th>Mean Scores</th>
<th>Differences of Means</th>
<th>t-value</th>
<th>P(t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 3 MEAP</td>
<td>2.28</td>
<td>.35</td>
<td>3.25</td>
<td>.002</td>
</tr>
<tr>
<td>AWC Holistic</td>
<td>2.63</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 4 MEAP</td>
<td>2.32</td>
<td>1.17</td>
<td>9.98</td>
<td>&lt;0.0005</td>
</tr>
<tr>
<td>AWC Holistic</td>
<td>3.49</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 5 MEAP</td>
<td>2.67</td>
<td>.14</td>
<td>1.52</td>
<td>.132</td>
</tr>
<tr>
<td>AWC Holistic</td>
<td>2.81</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In Table 5, Dependent Sample t-test of Mean Scores for MEAP and AWC, the mean scores of the AWC are higher for every grade level. At the grade 3 and 4 levels, the means are statistically different (.35 and 1.17, respectively) whereas the difference between the AWC and MEAP mean at Grade 5 is not statistically different. The t-value was calculated assuming a dependent sample. The probability associated with the t-value is less than half of a scale point at Grade 3 (.002) and at Grade 4 (<0.0005), the probability is more than a scale point difference.

Table 6. Correlation Study of MEAP and AWC.

<table>
<thead>
<tr>
<th></th>
<th>Holistic</th>
<th>Content</th>
<th>Structure</th>
<th>Stance</th>
<th>Sentence Fluency</th>
<th>Diction</th>
<th>Conventions</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 3</td>
<td>.53**</td>
<td>.57**</td>
<td>.49**</td>
<td>.57**</td>
<td>.59**</td>
<td>.49**</td>
<td>.28*</td>
</tr>
<tr>
<td>Grade 4</td>
<td>.56**</td>
<td>.63**</td>
<td>.54**</td>
<td>.65**</td>
<td>.57**</td>
<td>.58**</td>
<td>.38**</td>
</tr>
<tr>
<td>Grade 5</td>
<td>.70**</td>
<td>.70**</td>
<td>.70**</td>
<td>.64**</td>
<td>.67**</td>
<td>.67**</td>
<td>.53**</td>
</tr>
</tbody>
</table>

** Correlation is significant at the 0.01 level.
* Correlation is significant at the 0.05 level.

Correlations between the MEAP and AWC are at a significant degree at every grade level and attribute score. Of the seven scores available the MEAP correlates most highly with the AWC content, stance and sentence fluency scores. Significant 0.01 level for all bivariant correlations happen at all levels and attributes except for Grade 3 conventions which is significant at the 0.05 level. Correlations are moderate to strong with the highest correlations occurring with the holistic, content, stance, and sentence fluency. The lowest correlation is between the MEAP and conventions.

To determine the correlation (see table 7) between the comment codes and MEAP or AWC scores, the comment codes were coded 0 to mean absence of code and 1 for presence of code. Codes #1-4 have negative values suggesting an expected relationship between the comment code and the attribute. Code #1 has a clear relationship that is significant with
Table 7. Point Biserial Correlation between Comment Codes and MEAP or AWC.

<table>
<thead>
<tr>
<th>Comment Code</th>
<th>MEAP Rubric Score</th>
<th>AWC Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Holistic</td>
<td>Content</td>
</tr>
<tr>
<td>#1 Lacks focus on one central idea.</td>
<td>-.56**</td>
<td>-.28**</td>
</tr>
<tr>
<td>#2 Demonstrates limited control over sentence structure, vocabulary, and/or conventions.</td>
<td>-.03</td>
<td>-.08</td>
</tr>
<tr>
<td>#3 Needs details and examples to adequately develop the ideas and content.</td>
<td>-.19**</td>
<td>-.04</td>
</tr>
<tr>
<td>#4 Lacks coherent organization and/or connections between ideas.</td>
<td>-.08</td>
<td>-.10</td>
</tr>
<tr>
<td>#5 Needs richer development of central idea with some additional, relevant details and examples.</td>
<td>.53**</td>
<td>.25**</td>
</tr>
<tr>
<td>#6 Needs tighter control of organization and/or the connections among ideas.</td>
<td>.30**</td>
<td>.17*</td>
</tr>
<tr>
<td>#7 Need greater precision and maturity of language.</td>
<td>.39**</td>
<td>.26**</td>
</tr>
</tbody>
</table>

* Correlation is significant at the 0.05 level.
** Correlation is significant at the 0.01 level.
MEAP rubric (-.56), NWP holistic rubric (-.28), and the six AWC attributes: content (-.31), structure (-.27), stance (-.34), sentence fluency (-.31), diction (-.28), and conventions (-.23). Codes #2, 3, and 4 have eight different attributes for a total of 24 correlations. Of those correlations, only one, between comment code 3 and the MEAP rubric, is significant (-.19) and highly significant at that. Statistically codes #2, 3, and 4 have no relationship between the code presence and scores received. For codes #5, 6, and 7, all values fall in the positive range. The higher number represents a higher deficit between the comment code and the attributes.
Discussion

This analysis of the MEAP rubric and NWP’s AWC is based on papers scored by both scorers hired by the State of Michigan and the Upper Peninsula Writing Project TCs. The data supports the idea that the two rubrics are highly correlated. Yet, the MEAP scorers did not score any papers at a 5 or a 6 level on the MEAP rubric. However, the scorers using the AWC did score 22 papers at level 5 and 4 papers at level 6. This higher scoring led to higher mean scores for AWC especially at Grade 4 where a discrepancy occurred at the scoring levels of 1 and 2. The MEAP scorers gave 46 papers a 1 or a 2 whereas only 14 papers received the same scores from the AWC holistic scorers.

When considering the MEAP Comment Codes given by the MEAP scorers, over three-quarters of the papers received Comment Code #3 – needs details and examples to adequately develop the ideas and content. This comment code only correlates with the MEAP rubric score. But there is little correlation between code #3 and the AWC. The number of papers with a score of 1, 2, or 3 at Grade 3 was 71 samples with 63 of those papers receiving comment code #3. No distinction is made between the three levels of scoring by using the same comment code. This same pattern also occurs at Grades 4 and 5.

The second most commonly used comment code was # 4 – lacks coherent organization and/or connections between ideas and it has no relationship between the code presence and the score given by the scorers.

The distribution of the other six MEAP comment codes occurred on less than 12% of the scored papers. Comment code #8 - earned the highest score of 6 - was never used and if had been
used, would state the obvious and would be of no value to those interested in improving student writing.

The MEAP comment codes tend to mislead whereas AWC is correlated with MEAP scores. For learning from the results to occur, there must be something subtle in the distinctive difference between the comment codes and they must be internally consistent and exhaustive.

The AWC provides six different attributes and six levels for each attribute that teachers, administrators, and parents could consider in improving student writing. It would allow interested stakeholders to present lessons for improvements to students because it gave results on all six attributes whereas the MEAP rubric does not. This analysis provides support that the National Writing Project’s Analytic Writing Continuum [AWC] provides more consistent feedback to teachers, parents, and administrators.

Recommendations

Using the results of the analysis the following recommendations are made:

1. The State of Michigan’s Office of Educational Assessment and Accountability (OEAA) who administers the MEAP writing assessment should reconsider the use of the current comment codes as no relationship occurs between the comment code(s) given and the score received.

2. OEAA should revamp the current comment codes so that the codes better reflect why a paper receives the score from the scorer. OEAA may want to scrap the comment codes and consider an analytic rubric in addition to the holistic rubric currently used. The analytic rubric could be based on the 6 + 1 Traits (Culham, 2003).

3. School districts should request better feedback from the state to help them improve their students’ writing. The districts should also review the MEAP samples contained on the
MEAP Writing CD to verify the writing should receive the score reported. In reviewing the writing samples, the strengths and weaknesses of each writing sample will be observed and may bring clarity as to why the comment code was assigned to the writing. The comment code alone will not clarify the reason for the score received.

Areas of Further Research

As the above data is based on one school district’s writing assessments, further research should occur with the MEAP scores and comment codes to see if the same results occur across the State of Michigan. This exact study can not be replicated as only Grades 4 and 7 will be assessed on writing in the future. Field testing for the new writing assessment occurred in the fall of 2009 and Michigan will implement the new writing assessment in the autumn of 2010. Other changes to the test and the way results are reported may occur. However, the same study procedures can be used on future writing samples.

The subjects of a future study could be fourth and seventh graders from school districts randomly selected across Michigan. The writing samples would be randomly generated from the MEAP Writing CD of each district. The data would be collected on the MEAP scores and comment codes. Then teacher consultants (TCs) from several National Writing Project sites across Michigan would converge, be trained and calibrated in the use of the Analytic Writing Continuum (AWC). Then the TCs would score the writing samples using the AWC holistic rubric and the six attributes of the analytic rubric.

The data from the AWC scoring and the MEAP scoring would be first analyzed by looking at the characteristics of the students who wrote the papers and then the frequency distribution of the scores from both scoring procedures. The analysis should include the distribution of MEAP comment codes, the dependent sample t-test of mean scores, as well as the
correlation study of the MEAP and AWC, and the point biserial correlation between comment codes and MEAP or AWC. These analyses are the same that were conducted for this paper.

Conclusion

As state assessments continue to be important in the evaluation of school districts, the feedback from these assessments must provide information on how to improve student writing. The MEAP rubric and the NWP’s AWC scores from the same writing samples do correlate but the comment codes from the MEAP do not give school districts detailed information to implement improvements to the writing curriculum. Districts need to be informed about this weakness in the MEAP writing assessment results and must demand a change in the use of the comment codes.

Acknowledgements

The author is grateful to the Upper Peninsula Writing for providing the opportunity to write about the results of the scoring conference, Sherry Swain, Director Field Office, NWP, for her knowledge of the AWC, and to Paul LeMahieu, Director of Research and Evaluation, NWP for his assistance in the data and analysis.
References


Arter, J. & McTighe, J. *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*


doi:10.1016/s1075-2935(02)00029-6


doi:10.1016/j.edurev.2007.05.002


Kindergarten through grade 12. Retrieved August 4, 2009 from
http://www.michigan.gov/mde/0,1607,7-140-22709_31168-172665--,00.html

Considerations in Writing Assessment. Assessing Writing, 1, 147-173. doi:10.1016/1075-
2935(95)90021-7

College Composition and Communications, 50, 483-503. doi:10.2307/358862
Appendix A

Michigan Educational Assessment Program (MEAP)
Writing from Knowledge and Experience
Grades 3 - 8
Holistic Scorepoint Descriptions

6 The writing is exceptionally clear and focused. Ideas and content are thoroughly developed with relevant details and examples where appropriate. The writer’s control over organization and the connections between ideas moves the reader smoothly and naturally through the text. The writer shows a mature command of language including precise word choice that results in a compelling piece of writing. Tight control over language use and mastery of writing conventions contribute to the effect of the response.

5 The writing is clear and focused. Ideas and content are well developed with relevant details and examples where appropriate. The writer’s control over organization and the connections between ideas effectively moves the reader through the text. The writer shows a command of language including precise word choice. The language is well controlled, and occasional lapses in writing conventions are hardly noticeable.

4 The writing is generally clear and focused. Ideas and content are developed with relevant details and examples where appropriate, although there may be some unevenness. The response is generally coherent, and its organization is functional. The writer’s command of language, including word choice, supports meaning. Lapses in writing conventions are not distracting.

3 The writing is somewhat clear and focused. Ideas and content are developed with limited or partially successful use of examples and details. There may be evidence of an organizational structure, but it may be artificial or ineffective. Incomplete mastery over writing conventions and language use may interfere with meaning some of the time. Vocabulary may be basic.

2 The writing is only occasionally clear and focused. Ideas and content are underdeveloped. There may be little evidence of organizational structure. Vocabulary may be limited. Limited control over writing conventions may make the writing difficult to understand.

1 The writing is generally unclear and unfocused. Ideas and content are not developed or connected. There may be no noticeable organizational structure. Lack of control over writing conventions may make the writing difficult to understand.

Condition codes for unratable papers (zeroes):

A Off topic
B Written in a language other than English or illegible
C Blank or refused to respond

August 2005
Appendix B

Comment Codes: ELA Writing from Knowledge and Experience

1  Lacks focus on one central idea.

2  Demonstrates limited control over sentence structure, vocabulary and/or conventions.

3  Needs details and examples to adequately develop the ideas and content.

4  Lacks coherent organization and/or connections between ideas.

5  Needs richer development of the central idea with some additional, relevant details and examples to get a higher score.

6  Needs tighter control of organization and/or the connections among ideas to get a higher score.

7  Need greater precision and maturity of language use to get a higher score.

8  Earned the highest score of 6.

Source: Fall 2008 MEAP results