

CTB/McGraw-Hill

**West Virginia Writing Roadmap
2006 Validation Study Report**

Developed and published under contract with the West Virginia Department of Education by CTB/McGraw-Hill LLC, a subsidiary of the McGraw-Hill Companies, Inc., 20 Ryan Ranch Road, Monterey, California 93940-5703. Copyright © 2006 by the West Virginia Department of Education. Only State of West Virginia educators and citizens may copy, download and/or print the document, located online at <http://WESTEST.k12.wv.us>, for educational purposes only. Any other use or reproduction of this document, in whole or in part, requires written permission of the West Virginia Department of Education and the publisher.



Introduction

The purpose of the validation study is to obtain validity evidence on the automated scoring of West Virginia Writing Roadmap using the scoring engine. This algorithm is based on a mixture of artificial intelligence, natural language processing and statistical technologies (Elliot, 2001). The validity of automated scoring systems essentially rests on how well they can replicate the scores given by human raters. The typical validation study consists of two steps. The first step is to “train” the automated scoring algorithm by giving it a set of papers (e.g., 350) along with the numeric scores given by expert human raters. Once training has been completed for a given writing prompt using this information, the algorithm is ready to score validation papers. This step uses a second set of validation papers (e.g., 150) that were previously scored by raters. The algorithm is used to score this second set of “blind” papers. These papers are blind in that no numeric writing scores are included, just the students’ written responses. The scores from human raters are then compared with those generated from the scoring algorithm to determine how well they are in agreement. If high rates of agreement exist, then the claim is made that the algorithm has adequately captured human judgment. The simplest measure of rater agreement is “percent perfect agreement”. For example, perfect agreement exists when both the rater and scoring algorithm assign a “3” to a given paper. For a trait scored writing prompt, 40-70 percent perfect-agreement is considered to be acceptable. Additional insight on scoring accuracy can be obtained from other measures of rater agreement.

Rater Agreement

The overall results from the rater agreement analysis demonstrate very good rater agreements between the scoring engine and the human rater for trait scores from the additional 8 West Virginia (WV) prompts for the 2006 administration. WV students’ essay responses to the four Grade 7 writing prompts and four Grade 10 writing prompts are scored by both the engine and the CTB expert human raters in the 2006 validation study. The kappa statistic indicates the rater agreement beyond the chance level which is computed for WV Writing Road Map. Table 1 shows the weighted kappa statistics for trait scores for each of the 8 additional prompts.

Kappa statistics for both Grade 7 and Grade 10 prompts demonstrate very good consistency between human rater and the engine scores. Kappa statistics are in the .60 to .80 range for the trait scores overall. Kappa ranges from 0.70 to 0.87 for the four Grade 7 prompts and from 0.67 to 0.82 for the four Grade 10 prompts. These values of kappa statistic demonstrate good to excellent rater agreement beyond the chance agreement (Fleiss, J. Levin, B, and Paik, M., 2003). In the 2006 engine training and validation study, there is an increase in the overall number of training papers and blind validation papers as well as the number of papers at the extreme score points. As a result, it reduces the chance agreement and contributes to the improvement of the kappa statistics compared to that of the original 8 prompts validation study (Table 2).

Table 1 also presents the percent of perfect agreement and the percent of adjacent agreement (cumulative) for these 8 prompts. The perfect agreement rates are in the acceptable range of 40%-60% for all prompts. The adjacent agreement rates for all prompts are at and above 90% with the lowest being 90% and 91% for 10D1 Mechanics and Sentence Structure. The discrepant rates for 10D1 Mechanics and Sentence Structure are around 10%. This may be caused by the lacking of examples of the students’ writing related to these two traits in the training papers.



Reference

- Elliot, S. (2001). Intellimetric™: From Here to Validity. In *Automated Essay Scoring: A Cross Disciplinary Perspective* Shermis, M.D. & Burstein, J (Eds.) Lawrence-Erlbaum, Mahwah: NJ.
- Fleiss, J. L. Levin, B. Paik, M.C. (2003). *Statistical Methods for Raters and Proportions* (Third Edition). New York: John Wiley & Sons, Inc.

Table 1. 2006 WV WRM Validation Study

Form (Number of Papers)	Trait	Consistency	Percent of Agreement	
		Kappa	Perfect	Sum of Perfect and Adjacent
7D2 (167)	Organization	0.85	56	97
	Development	0.82	41	98
	Sentence structure	0.84	56	96
	Word choice	0.79	48	95
	Mechanics	0.82	52	96
7E2 (169)	Organization	0.84	54	95
	Development	0.84	57	95
	Sentence structure	0.85	57	97
	Word choice	0.84	54	98
	Mechanics	0.85	54	97
7N2 (159)	Organization	0.83	50	96
	Development	0.87	65	97
	Sentence structure	0.78	42	94
	Word choice	0.71	41	94
	Mechanics	0.76	43	95
7P2 (209)	Organization	0.70	50	93
	Development	0.70	45	94
	Sentence structure	0.77	56	95
	Word choice	0.73	56	94
	Mechanics	0.74	55	96
10D1 (223)	Organization	0.74	47	93
	Development	0.79	47	96
	Sentence structure	0.71	45	91
	Word choice	0.76	49	95
	Mechanics	0.73	49	90
10E1 (256)	Organization	0.76	48	93
	Development	0.79	50	94
	Sentence structure	0.78	52	94
	Word choice	0.77	50	95
	Mechanics	0.78	48	96
10N1 (150)	Organization	0.81	59	96
	Development	0.81	60	97
	Sentence structure	0.75	51	94
	Word choice	0.78	53	97
	Mechanics	0.74	49	93
10P2 (208)	Organization	0.72	55	97
	Development	0.82	60	100
	Sentence structure	0.72	54	97
	Word choice	0.67	51	96
	Mechanics	0.72	51	97

Table 2. 2005 WV WRM Validation Study

Form (Number of Papers)	Trait	Consistency	Percent of Agreement	
		Kappa	Perfect	Sum of Perfect and Adjacent
7D1 (150)	Organization	.55	49	94
	Development	.71	53	96
	Sentence structure	.68	53	97
	Word choice	.57	58	97
	Mechanics	.59	49	97
7E1 (193)	Organization	.47	47	95
	Development	.64	58	96
	Sentence structure	.54	43	96
	Word choice	.51	57	96
	Mechanics	.50	47	95
7N1 (111)	Organization	.74	55	96
	Development	.82	61	100
	Sentence structure	.63	54	99
	Word choice	.60	46	98
	Mechanics	.68	56	98
7P1 (140)	Organization	.26	38	96
	Development	.43	48	97
	Sentence structure	.31	49	95
	Word choice	.39	61	98
	Mechanics	.47	54	99
10D2 (145)	Organization	.70	54	94
	Development	.73	61	94
	Sentence structure	.69	51	92
	Word choice	.64	52	92
	Mechanics	.60	52	88
10E2 (190)	Organization	.68	43	92
	Development	.65	40	91
	Sentence structure	.53	40	89
	Word choice	.61	46	93
	Mechanics	.64	50	95
10N2 (168)	Organization	.65	60	95
	Development	.70	59	98
	Sentence structure	.63	49	96
	Word choice	.69	55	98
	Mechanics	.70	54	98
10P1 (137)	Organization	.66	53	96
	Development	.73	53	96
	Sentence structure	.69	48	97
	Word choice	.64	53	98
	Mechanics	.73	59	97