The McGraw-Hill Companies

CTB McGraw-Hill

CTB/McGRAW-HILL

West Virginia Writing Roadmap Validation Study Report

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 8 West Virginia (WV) prompts. Four Grade 7 writing prompts and four Grade 10 writing prompts were scored by both the engine and the CTB human raters in the 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 and the absolute percent agreement for each trait score from each prompt.

Kappa statistics for Grade 10 prompts demonstrate very good consistency between human rater and the engine scores. Kappa statistics are in .60 to .70 range (except .53 for 10E2 Sentence Structure). Kappa ranges from 0.47 to 0.82 for three of the Grade 7 prompts with Kappa in .50-.60 range for most of trait scores. These statistics demonstrate good rater agreement. For Grade 7 Persuasive Prompt 1 (7P1), the kappa for the trait scores Organization and Sentence Structure are .26 and .31, below the acceptable range. The rubric improvement is recommended in order to further clarify the assignment of scores.

Table 1 also presents the percent of perfect agreement and % of adjacent agreement (cumulative) for these 8 prompts. The perfect agreement rates are in acceptable range of 40%-60% for all prompts except for 7P1 trait 1 Organization (38%). Although Organization for 7P1 trait 1’s perfect agreement rate is 38%, slightly below the acceptable range, the cumulative adjacent agreement is 96%. The adjacent agreement rates for all prompts are at and above 90% except for one trait score of 88%. The discrepant rate for 10D2 Mechanics is 12% and the rubric should be examined.

Reference

Elliot, S. (2001). Intellimetric™: From Here to Validity. In A*utomated Essay Scoring; A Cross Disciplinary Perspective* Shermis, M.D. & Burstein, J (Eds.) Lawrence-Erlbaum, Mahwah: NJ.

Table 1

Form (Number of Papers)	Trait	Consistency	Percent of Agreement	
		Kappa	Perfect	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