Assessing Writing 21 (2014) 89–103

Contents lists available at ScienceDirect

Assessing Writing

The *WPA Outcomes Statement*, validation, and the pursuit of localism

Diane Kelly-Riley[a],*, Norbert Elliot[b],**

[a] University of Idaho, Moscow, ID 83844, USA
[b] New Jersey Institute of Technology, University Heights, Newark, NJ 07102, USA

ARTICLE INFO

Article history:
Received 14 October 2013
Received in revised form 13 March 2014
Accepted 15 March 2014

Keywords:
Common European Framework of Reference for Languages
Consensus statements
Localism
Writing assessment
Validation
WPA Outcomes Statement for First-Year Composition

ABSTRACT

This validation study examines the *WPA Outcomes Statement for First-Year Composition*, a United States consensus statement for first-year post-secondary writing, as implemented in a unified instructional and assessment environment for first-year college students across three different institution types. Adapting categories of contemporary validation from Kane (2013), we focus on four forms of evidence gathered from early and late-semester student performance ($n = 153$): scoring, generalization, extrapolation, and implication. With an emphasis on education policies in action, the study generates important questions that, in turn, provide a basic framework for further research into the challenges of aligning broad consensus statements with locally developed educational initiatives.

> To validate an interpretation or use of test scores is to evaluate the plausibility of the claims based on the scores. An argument-based approach to validation suggests that the claims based on the test scores be outlined as an argument that specifies the inferences and supporting assumptions needed to get from test responses to score-based interpretations and uses. Validation then can

* Corresponding author at: Department of English, University of Idaho, 875 Perimeter Drive, MS 1102, Moscow, ID 83844-1102, USA. Tel.: +1 208 885 5704.
** Corresponding author at: Department of Humanities, New Jersey Institute of Technology, University Heights, Newark, NJ 07102, USA. Tel.: +1 973 596 3266.
E-mail addresses: dianek@uidaho.edu (D. Kelly-Riley), elliot@njit.edu (N. Elliot).

http://dx.doi.org/10.1016/j.asw.2014.03.004
1075-2935/© 2014 Elsevier Ltd. All rights reserved.

be thought of as an evaluation of the coherence and completeness of this interpretation/use argument and of the plausibility of its inferences and assumptions. (Kane, 2013, p. 1).

1. Introduction

In the United States, the *WPA Outcomes Statement for First-Year Composition* has been designed to identify common knowledge, skills, and attitudes desired by post-secondary entry level composition programs (Council of Writing Program Administrators, 2000/2008). Intended to introduce first-year students to writing expectations in post-secondary settings, the *Outcomes Statement* provides a common administrative and instructional configuration for American higher education. The *Outcomes Statement* is a consensus statement developed, amended, and used by the Council of Writing Program Administrators, a non-profit organization organized for educational and scientific purposes (Harrington, Malencyzk, Peckham, Rhodes, & Yancey, 2001; O'Neill, Adler-Kassner, Fleischer, & Hall, 2012). With its five outcomes (rhetorical knowledge; critical thinking, reading, and writing; processes; knowledge of conventions; and composing in electronic environments) and twenty five traits supporting them, the *Outcomes Statement* has been used as the basis for assessment of student writing and for the evaluation of program outcomes. The *Outcomes Statement* "articulates what composition teachers nationwide have learned from practice, theory and research" (Thomas, 2013, p. 165).

The impetus for the *Outcomes Statement* came in response to pressures from the six United States regional education accrediting agencies—Middle States Association of Colleges and Schools Commission on Higher Education; New England Association of Colleges and Schools Commission on Institutions of Higher Education; North Central Association of Colleges and Schools Higher Learning Commission; Northwest Commission on Colleges and Universities; Southern Association of Colleges and Schools Commission on Colleges; and Western Association of Colleges and Schools Accrediting Commission of Senior Colleges and Universities—for institutions to demonstrate student learning outcomes embedded within programmatic and institutional settings. While the U.S. Department of Education does not accredit institutions or their programs, the Secretary of Education is required by law to authenticate these agencies as authorities capable of evaluating the quality of education in programs they accredit. Because they are responsible for educational quality, these accrediting agencies have a great deal of leverage in requiring institutions to demonstrate that learning outcomes are used to improve student performance on specific educational domains such as writing ability.

The ever present American tension between federalism (the role of a strong central government) and localism (the resistance against such a government in favor of regional autonomy) is expressed by the original contributors to the *Outcomes Statement*. As Rhodes, Peckham, Bergmann, and Condon (2005) observed, "We confronted an unpleasant fact: the term *first-year composition* varied widely in meaning. . .the term was hotly contested among the very people in charge of administering it. So we asked ourselves, if we couldn't agree what first-year composition should be, how could we ever account for what we do?" (p. 12). Operationally, it was difficult for Writing Program Administrators (WPAs) to reconcile their particular state and institutional contexts—institutional missions, the demographic make-up of their students, instructional faculty, the configurations of their departments—with national consensus statements specifying what students should learn in their first year of college-level writing.

Resolution of these tensions was found in the emphasis on broad outcomes—"what students exiting first-year composition should know and be able to do" (Rhodes et al., 2005, p. 12)—not on standards and accompanying levels of performance. According to Yancey (2005), the focus on outcomes was a way to recognize the unique local situation of first-year composition programs while providing curricular stability that resonated on a national level. "While outcomes articulate the curriculum, they do not specify how well students should know or understand or do what the curriculum intends. . .Because outcomes are not benchmarked against levels of performance, individual programs or institutions can have the same curricular outcomes but have different ideas about when and how well they want students to perform" (Yancey, 2005, p. 22). As a result of the focus on outcomes and the absence of levels of performance, a first-year writing program at a two-year, rural community college focused on retraining adult, displaced factory workers in the state of Michigan could have the same outcomes as undergraduates enrolled in the highly-selective first-year writing program at Harvard University. Recent publications (Behm, Glau, Holdstein, Roen, & White, 2013; Harrington, Rhodes, Fischer, & Malencyzk, 2005) detail the origin for the collaboration that defined the *Outcomes Statement*, provide historical overview, analyze the outcomes themselves, and establish curricular applications in particular settings.

Nevertheless, as Isaacs and Knight (2013) have demonstrated, the *Outcomes Statement* has not been broadly adopted by American post-secondary universities. Three reasons are apparent in the description of the project provided above.

First, as a consensus document, the *Outcomes Statement* does not have the validity force of a reflective latent variable model. As Pellegrino and Hilton (2012) have defined this model in their recent National Research Council report, a reflective latent variable model is an empirical investigation, often using factor analysis, used to identify relationships among scores that, in turn, point to variables that appear to be the hidden cause of the performances of study participants. Identification of the Big Five personality factors, for example, is the result of a series of twentieth-century statistical studies that have identified five dimensions of personality: openness to experience; conscientiousness; extraversion; agreeableness; and emotional stability (Goldberg, 1993). Conversely, consensus statements such as the *Outcomes Statement* in the United States and the Common European Framework of Reference for Languages (CEFR)—published in 2001 to establish language qualifications in the service of educational and occupational mobility—are formative latent variable models (Council of Europe, 2001). Distinct from the formal methods that have characterized taxonomies such as the Big Five personality factors, these formative, emerging models are established through consensus of experts, supported by literature review and tradition. As such, longitudinal studies of empirical research do not play the critical role in formative models as they do in reflective models—and thus accompanying empirical validity arguments are often sacrificed in favor of claims that consensus has been reached.

Second, as these claims become the center of debate, formative models are often left in the hands of small groups of practitioners for implementation. In the United States, these are the Writing Program Administrators (WPAs). As Malenczyk (2013) has defined this group, WPAs who began their careers in the mid-to-late 1990s may have been appointed as the administrator responsible for a post-secondary writing program merely because they had an interest in the field of writing studies. While those just beginning their careers with WPA appointments have distinctly more career investment in the success of a writing program (Charlton, Charlton, Graban, Ryan, & Stolley, 2011), the consistency of faculty development programs contributing to the failure or success of the *Outcomes Statement* use at a specific institutional site remains unknown at best and uneven at worst. Such implementation complexities have also been demonstrated in terms of the CEFR (Neff-van Aertselaer, 2013).

While the founders of the *Outcomes Statement* wisely declined to establish performance levels for the increasingly diverse post-secondary American educational system, they also did not establish a validation system for the five outcomes and twenty five supporting traits that could be used by local institutions. This absence of local validation processes constitutes the third reason for the apparent failure of adoption. In similar fashion, as Little (2007) has claimed in the case of CEFR impact, challenges remain in terms of lasting curricular impact. Absent in both America and Europe appears to be a system by which consensus statements can be assessed, refined, and implemented in terms of local curricula, instructors, and students. In the case of the *Outcomes Statement*, a system of validation can be used to examine student scores on the five outcomes—scores that can, in turn, be used to establish a series of interpretative arguments supporting broad-based use in local settings and across differing institution types. The system of validation can serve as a vehicle to ensure that consensus statements are examined, tailored, and used to benefit students.

As a needed logical step in validating and advancing the use of this avant-garde consensus statement, this descriptive, exploratory study examines the empirical qualities of the *WPA Outcomes Statement* as assessed in pedagogical settings across three first-year writing programs at three different institution types. These writing programs used the *Outcomes Statement* to support instruction and define assessment within a cohesive, student-centered learning environment. Adopting the validation framework provided by Kane (2013), we focus on four types of validation evidence: scoring, generalization, extrapolation, and implication. While the sample is relatively small ($n = 153$) and confined to the participating institutions, the study nevertheless generates important questions that, in turn, provide a starting point for further research as educational institutions in the United States and Europe work to implement consensus statements in ways that advance learners' language use in the age of globalization.

2. Validation framework

The *Standards for Educational and Psychological Testing* established a re-conceptualized definition of validity that prioritized the use and interpretation of test scores in a particular setting as the most fundamental consideration related to measurement (AERA, NCME, & APA, 1999). These *Standards* revised the long-standing view of validity as component categories—construct, content, predictive validity—and held reliability as a distinct category of evidence equally important to validity (Brennan, 2006). The 1999 revision of the construct of validity dramatically changed the paradigm for what underlined good practice in educational measurement. The enemy of assessment became construct underrepresentation. Many standardized tests have therefore come to be seen as limited because, in their insistence on privileging one category of validity—such as score reliability—over others, they have not robustly engaged the cognitive, intrapersonal, and interpersonal domains of the construct under examination. A refreshed concept of validity from the field of educational measurement—a process of gathering evidence that, in turn, can be used to justify assessment use—informs the research we report in this paper.

While the *Standards* are under revision (Camara & Ernesto, 2011), the process of validating the interpretation and uses of test scores has been expanded and deepened by Kane (2006, 2013). Based on an assessment interpretation/use argument (IUA) in which result are interpreted and audiences are persuaded based on the plausibility of claims, Kane identifies sources of validity evidence in terms of four sources: scoring, defined as the appropriateness and application of evaluative rules; generalization, defined as the relationship of the observed score on a particular assessment to its limited representation in target domain; extrapolation, defined as the extension of this limited representation to the full range of performances in the target domain; and implication, defined as the extended interpretation and use of results. An IUA developed in association with the *Outcomes Statement* would thus broadly address the four categories of validity evidence in the following way: scoring procedures would be developed which allowed a range of scores (appropriateness) that readers could use (application) to evaluate the papers; generalization data would be gathered to establish a claim that the scores reflect the five outcomes identified in the *Outcomes Statement*; extrapolation data would be gathered to extend the *Outcomes Statement* to the full range of writing performances in the target domain; and implications would be drawn regarding the interpretation and use of the information derived in the validation effort.

While there are many kinds of processes, analyses, and arguments to be made for each of these four sources of evidence, IUA provides a validation process that combines and manages the evidence sources for the study at hand. This sense of precision can also be used to identify that which is being validated: score use. Because this is a descriptive, exploratory case study designed to examine the applicability of IUA to validate a particular consensus statement, the scores earned by students may be taken as evidence that the variables of the *Outcomes Statement* can be used in a local setting to design a curriculum amenable to assessment. In its fundamental orientation, the present study follows the evidence-centered design (ECD) framework of Mislevy, Steinberg, and Almond (2002). As an explicitly criterion-based model, ECD is intended for assessments using complex tasks in contexts where standardized, off-the-shelf assessments simply will not do because of their slender representations of the construct under examination.

As shown in the following section, in the ECD framework that is implicit in this study, participating instructors knew that the central instructional and assessment structure was based on the *Outcomes Statement*. As such, the outcomes are understood, from the first steps of curricular design, to be the variables of the study that were to be used to both frame the curriculum and assess the ways that students used these variables. In other words, the *Outcomes Statement* served as a heuristic to define the construct of writing for the project, and then specific variables were later implemented as the assessment criteria for the project. As a representation of the force of the *Outcome Statement* to create evidentiary connections between instruction and assessment, test scores are thus at the center of the study.

3. Study context

The study finds its origin in the Southern California Outcomes Research in English (SCORE), sponsored by the McGraw-Hill Companies. SCORE explored a common set of outcomes (a rubric with variables drawn from the *Outcomes Statement*) and supporting technology (a software platform used for course management) that could be used across different first year composition programs in a variety of institutional settings to support a uniform framework for instruction and assessment purposes. (Similar ECD project methodology has been described in Condon & Kelly-Riley, 2004.) In 2010–2011, the publishing company McGraw-Hill sponsored a project across three post-secondary institution types—a community college, a small, private liberal arts college and a public university located in Southern California—using its electronic textbook and course management platform, *Connect*, in conjunction with *The McGraw-Hill Handbook* (Maimon, Peritz, & Yancey, 2011). The three writing program administrators collaborated to adapt the *Outcomes Statement* for use by instructors at the three institutions to serve as the basis for instruction and to respond to students' work in the classroom.

During the academic terms, the three institutionally-based WPAs worked with instructors to tailor the *WPA Outcomes Statement* to their particular instructional settings. This process of localization allowed each institution to leverage the advantages of a common variable framework for instruction and assessment while adapting those variables in ways meaningful to the local needs of students. In reality, the *Outcomes Statement* emerged as a heuristic allowing WPAs and instructors to plan their classroom activities, structure the types of writing assignments students would do, and serve as a formative feedback vehicle for response to student writing. The *Outcomes Statement* functioned as Wiley (2005) describes: "general goals for writing programs that can serve as a heuristic for designing various curricula and pedagogies whose ends are similar, but at the same time vary in form and content, emphasis and sequence" (p. 29). At the end of the curricular design process, WPAs identified four outcomes for assessment and instruction within their writing programs based on the *Outcomes Statement*—rhetorical knowledge; evidence; structure; and language. The collaborators made decisions about the variables most important to them and reduced the 25 traits on the *WPA Outcomes Statement* to reflect twelve that they wanted to emphasize in their programs. The SCORE collaborators did not include the assessment of writing process or composing in electronic environments. Table 1 maps the broad outcomes and their associated traits of the *Outcomes Statement* to the SCORE study.

To unify instruction with assessment, the three faculty WPAs from these institutions agreed upon requirements for two common analytically-based writing assignments. The first assignment—given at the beginning of the term—was 300–500 words in length and required to students to use a single outside source. Such an assignment reflected the kinds of diagnostic writing tasks that are often used early in the course in American post-secondary institutions to provide a rapid assessment of writing ability. The second assignment—given at the end of the term—was a 5–7 page argumentative essay that required students to synthesize several outside sources. This kind of assignment reflected the kinds of summative writing tasks that are often used later in a term to provide a robust assessment of writing ability. The writing assignments were adapted for the first year writing program contexts at the community college (a two year school), the public-research university, and the private liberal arts college. All of the participating institutions agreed to assign the two common writing assignments at the beginning and end of the terms. As constructed response assignments (Bennett, 1993), these writing tasks were designed to allow a twelve-dimension rubric—the eleven traits identified in Table 1 with an additional holistic score—to provide feedback to their students on the two common assignments. For each of the traits, readers awarded scores on a traditional six-point scale, with the score of 1 established as the low score and a score of 6 established as the high score. Papers were written in a naturalistic setting in which students could work on their assignments away from class, and students had the opportunity for process writing by receiving peer and instructor feedback on drafts of their writing assignments. Because this was a descriptive, base-line study

Table 1
WPA outcomes mapped to the SCORE outcomes.

WPA broad outcome	WPA specific trait	SCORE broad outcome	SCORE specific trait
Rhetorical knowledge	Focus on a purpose	Rhetorical knowledge	Thesis
Rhetorical knowledge	Adopt appropriate voice, tone and level of formality	Rhetorical knowledge	Expertise
Rhetorical knowledge	Use conventions of format and structure appropriate to the rhetorical situation	Rhetorical knowledge	Genre
Critical thinking, reading, and writing	Integrate their own ideas with those of others	Evidence	Ideas
Rhetorical knowledge	Respond to the needs of different audiences	Evidence	Credibility
Critical thinking, reading, and writing	Evaluating, analyzing, and synthesizing appropriate and secondary sources	Evidence	Integration
Rhetorical knowledge	Focus on a purpose	Structure	Focus
Knowledge of conventions	Develop knowledge of genre conventions. . .paragraphing	Structure	Paragraphs
Knowledge of conventions	Control such surface features as syntax, grammar, punctuation and spelling	Language	Correctness
Knowledge of conventions	Develop knowledge of genre conventions tone	Language	Eloquence
Knowledge of conventions	Practice appropriate means of documenting their work	Language	Documentation

to investigate student scores on the two assessments as evidence of the capability of the *Outcomes Statement* to provide a cohesive framework for instruction and assessment, the naturalistic setting was not intended to pair pre-and-post instructional writing tasks or establish treatment and control groups.

While there were 332 students enrolled in these courses, resources did not allow each early and late semester paper to be read. Time, finances and personnel limitations resulted in a smaller sample size being selected for this study. A randomly selected sample—constituting 48% of the sample ($n = 332$)—were taken for the project, representing students who had submitted a paper from early in the semester ($n = 153$, before substantial instruction was completed) and one later in the semester ($n = 153$, when instruction was completed). Early and late-semester papers were read at a single session by 24 instructors at the semester's end, and instructors did not read their own student's papers. Calibration was achieved through a fixed procedure. Before the rating session, anchor essays and ratings were discussed with all of the raters; raters then read at a variable speed and were not required to meet a certain quota; and scores that differed by more than one point (e.g., 6 and 4) were resolved by a third reading (O'Neill et al., 2009).

To establish inter-reader agreement and inter-reader reliability, 30% ($n = 46$) of the papers were read twice. Because each of the 46 papers was read twice, the total range of combined scores was 2 (low) to 12 (high). These papers were randomly selected from the sample of 306 essays. To present scores in a uniform way, scores on papers that were not read twice were simply doubled so that the range of scores of the papers read twice (2–12) was used for all 153 scores. While such a pattern of score reporting would not be acceptable for assessments such as student placement, course completion, or curricular advancement, our study did not seek to use scores for such high stakes purposes.

Additionally, we used students' SAT writing scores, as well as course grades in their first-year writing courses in which the curriculum had been embedded, to establish sources of extrapolation evidence.

4. Research questions

Based on the application of the Interpretation of Use Argument (IUA) to the tailored traits of the *Outcomes Statement* operationalized in the study context, our research focuses on the following assessment questions:

1. What do patterns of inter-reader agreement and inter-reader reliability reveal regarding the ability of the *Outcomes Statement* to provide a set of variables that may approximate known consensus and consistency estimates?
2. What do patterns of correlation and regression reveal regarding generalization evidence that the traits of the construct model are related?
3. What do patterns of correlation reveal regarding extrapolation evidence of the relationships among the traits to the external criterion measures such as the SAT Writing score and the course grade?
4. What do patterns of score increase reveal regarding the implications of using the *Outcomes Statement* as an integrated instructional and assessment framework?

5. Results

Results are presented in terms of our four categories of validity evidence.

5.1. Scoring

For scoring, we examined reliability measures of consensus and consistency (Stemler, 2004). First, patterns of agreement were examined between the raters according to the rubric items on the early and late papers. Second, a more rigorous form of reliability is reported—non adjudicated Pearson; the adjudicated Pearson. Stemler (2004) argues that these consensus and consistency measures are important in establishing score reliability but concedes that calculating and interpretation of these measures can be difficult. In cases of interpretative complexity, we hold that data disaggregation is key to the process. We have followed such a principle in our scoring analysis.

Table 2 includes detail about the agreement by raters on each rubric item by the proximity of the scores assigned by the raters. The consensus indicators reflect one of four categories: exact agreement when raters agreed on the same score; adjacent when raters agreed by one score (either more or less than the other rater's score); and then scores that represent differences by two or three points. Scores on student papers that were either exact or adjacent required no second reading; scores beyond adjacency required a third reading to resolve the discrepant score. Combined scores from the two raters ranged from 2 (low) to 12 (high).

Table 2 shows how the individual variables were evaluated in the first and second papers. Largely, the raters employed all or most of the entire range of scores for each rubric item.

In the early term papers exact and adjacent scores—those requiring no adjudication—ranged from 74% on expertise, ideas, paragraphs, and documentation traits to 90% on eloquence. In the later papers, papers requiring no adjudication ranged from 49% on the paragraph trait to 89% on correctness. There were lower levels of agreement on all scores for the late term papers. Overall, 82% of the papers from early in the term were in exact or adjacent agreement, and 76% of papers from late in the term were in exact or adjacent agreement. With the exception of the paragraph trait in the late term assessment, these rates of inter-reader agreement are similar to those reported in Collins, Elliot, Klobucar, and Deek (2013) in which 598 papers of first-year college students scored according to *Outcomes Statement* trait models yielded exact plus adjacent ranges from 70% to 98%.

Table 2 also provides consistency evidence as measured by the Pearson product moment correlation. Papers that were adjacent were not adjudicated; however, papers that differed by more than one point were read by a third reader. To foster transparency in the reading process, we report both correlations on both papers needing no adjudication and those that did. The early term non-adjudicated Pearson scores range from .3 ($p < .05$) to .56 ($p < .01$). The late non-adjudicated Pearson scores fall in from .11 (*ns*) to .52 ($p < .01$). Early term adjudicated scores ranged from .72 ($p < .01$) to .89 ($p < .01$), and late term adjudicated scores ranged from .75 ($p < .01$) to .87 ($p < .01$). These rates of inter-reader reliability are also similar to—and, in some cases, stronger than—those reported in Collins et al. (2013). In that study, non-adjudicated papers scored according to a trait model ranged from .29 (ns) to 72 ($p < .01$). Adjudicated scores ranged from .48 ($p < .01$) to .86 ($p < 01$).

Both the early and late term adjudicated Pearson scores of the present study fall within the .7 range, understood as the threshold score for inter-reader reliability (Stemler, 2004; Williamson, Xi, & Bryer,

Table 2
Consensus and consistency estimates, early and late term papers ($n = 46$).

	Consensus estimates				Consistency estimates	
	Exact agreement	Adjacent	Scores differ by 2	Scores differ by 3	Non adjudicated Pearson	Adjudicated Pearson
Writing sample 1						
1. Thesis	17 (36%)	19 (40%)	5 (11%)	6 (13%)	.34*	.83**
2. Expertise	18 (38%)	17 (36%)	10 (21%)	2 (4%)	.42**	.78**
3. Genre	19 (40%)	21 (45%)	4 (9%)	1 (2%)	.56**	.72**
4. Ideas	16 (34%)	19 (40%)	9 (19%)	3 (6%)	.37*	.8**
5. Credibility	13 (28%)	26 (55%)	6 (13%)	2 (4%)	.38**	.75**
6. Integration	18 (38%)	21 (45%)	5 (11%)	3 (6%)	.3*	.73**
7. Focus	11 (23%)	25 (53%)	7 (15%)	4 (9%)	.42**	.79**
8. Paragraphs	17 (36%)	18 (38%)	9 (19%)	3 (6%)	.42**	.77**
9. Correctness	20 (43%)	22 (47%)	2 (4%)	3 (6%)	.53**	.79**
10. Eloquence	15 (32%)	26 (55%)	5 (11%)	1 (2%)	.5**	.73**
11. Documentation	19 (41%)	15 (33%)	7 (15%)	1 (2%)	.53**	.89**
12. Holistic	16 (35%)	22 (48%)	6 (13%)	2 (4%)	.5**	.76**
Writing sample 2						
1. Thesis	22 (47%)	11 (23%)	7 (15%)	7 (15%)	.42**	.87**
2. Expertise	23 (50%)	12 (26%)	8 (17%)	3 (7%)	.51**	.85**
3. Genre	19 (40%)	20 (43%)	5 (11%)	3 (6%)	.45**	.8**
4. Ideas	19 (40%)	16 (34%)	8 (17%)	4 (9%)	.39**	.79**
5. Credibility	14 (30%)	20 (43%)	7 (15%)	5 (11%)	.26 *ns*	.78**
6. Integration	10 (21%)	25 (53%)	8 (17%)	4 (9%)	.45**	.8**
7. Focus	19 (40%)	16 (34%)	6 (13%)	6 (13%)	.38**	.8**
8. Paragraphs	13 (8%)	19 (41%)	9 (20%)	5 (11%)	.11 *ns*	.76**
9. Correctness	17 (37%)	24 (52%)	3 (7%)	2 (4%)	.52**	.77**
10. Eloquence	19 (41%)	19 (41%)	5 (11%)	3 (7%)	.46**	.78**
11. Documentation	15 (32%)	15 (32%)	12 (26%)	5 (11%)	.38**	.75**
12. Holistic	16 (40%)	14 (35%)	6 (15%)	4 (10%)	.45**	.83**

* $p < .05$.

** $p < .01$.

2012). In this initial phase of the study, acceptable levels of inter-reader reliability had been achieved in the adjudicated scores for the purpose of examining the variables of the *Outcomes Statement* to be scored reliably, and these rates of inter-reader reliability were statistically significant. As a result, the adjudicated scores were thus appropriate to be used for the next phases of the study: generalization, extrapolation, and inference analysis.

5.2. Generalization

For evidence used to support the generalization inference that the traits of the construct model are related, we examine how well the construct of "good writing" defined by the traits is correlated in the early and late term papers. Since the *Outcomes Statement* is based on variables that may or may not be related, demonstrations of variable relationship is important. Those relationships are reported in Table 3.

In the early papers, all variables correlated according to the Pearson correlation with the range of .35–.85 ($p < .01$). High correlations are shown between the traits and the holistic scores, ranging from .51 to .83. In the later term papers, all variables correlated with the range of .44 to .87 ($p < .01$). The highest correlations again occur again between the holistic scores and the individual variables, ranging from .61 to .87. The high statistical significance and strength of the relationships among the variables provide evidence that the traits—mapped from the *Outcomes Statement* to SCORE study in Table 1—are related to each other.

Employing the eleven traits as the independent (or predictor) variables and the holistic score as the dependent (outcome) variable, regression analysis further demonstrates the strength of the model. For the early term paper, 88% of writing quality was accounted for by the ability of the traits to predict the holistic score: $R^2 = .89$, F (11, 146) = 95.2, $p < .01$. In the later term paper, 93% of writing quality was accounted for by the ability of the traits to predict the holistic score: $R^2 = .93$, $F(11, 121) = 135.28$, $p < .01$. Both of these analyses demonstrate excellent model strength.

5.3. Extrapolation

Table 4 presents late semester correlations between model established through generalization inference—that is, the 11 trait scores and the holistic score—and two criterion measures: SAT Writing scores as a measure of writing ability taken before the term began and the course grade assigned at the conclusion of the course. Grades ($n = 53$) were used from only one institution.

A statistically significant relationship is found between the SAT Writing and holistic Score: $r = .32$, $p < .05$. Similarly, a statistically significant relationship is found between the SAT Writing score and the course grade: $r = .46$, $p < .01$.

Lack of statistical relationship, however, was found between holistic score and course grade: $r = .23$ (*ns*). Indeed, only the traits of thesis, credibility, and eloquence correlated at statistically significant levels to the course grade. Subsequent analysis revealed the reason for the lack of relationship: the course grades lacked distribution. Specifically, only one grade of D was assigned, while 30 grades of B were assigned and 12 grades of A. Because the trait scores for these students demonstrated a full range of scores (2, 12) with a range of means (8.26–9.36), the lack of statistically significant relationship between trait scores and course grade is due to lack of normal distribution in the course grades.

5.4. Implication

Table 5 presents what many assessment stakeholders would consider the most important consequence of the study: statistically significant score increases across all variables. These statistically significant increases suggest that the *Outcomes Statement* can indeed be used to support instruction and define assessment within a cohesive, student-centered learning environment.

Table 5 details a paired sample t-test of mean scores for each of the variables for early and late course measures, and each variable demonstrates statistically significant increases in mean scores. These variables are also the same used in Table 2.

Table 3
Summary of early and late semester correlations ($n = 153$).

Early semester correlations	1	2	3	4	5	6	7	8	9	10	11	12
1. RK: Thesis	–	.57**	.65**	.78**	.48**	.7**	.74**	.66**	.56**	.6**	.44**	.83**
2. RK: Expertise		–	.66**	.57**	.56**	.53**	.7**	.61**	.64**	.66**	.31**	.72**
3. RK: Genre			–	.68**	.53**	.64**	.74**	.69**	.63**	.71**	.43**	.77**
4. Evidence: Ideas				–	.48**	.76**	.74**	.63**	.56**	.61**	.46**	.81**
5. Evidence: Credibility					–	.56**	.63**	.51**	.48**	.54**	.47**	.64**
6. Evidence: Integration						–	.66**	.54**	.5**	.56**	.57**	.78**
7. Structure: Focus							–	.78**	.64**	.67**	.43**	.83**
8. Structure: Paragraphs								–	.57**	.62**	.35**	.74**
9. Language: Correctness									–	.85**	.39**	.68**
10. Language: Eloquence										–	.47**	.73**
11. Language: Document											–	.51**
12. Holistic score												–
Late semester correlations												
1. RK: Thesis	–	.7**	.74**	.75**	.62**	.68**	.75**	.68**	.5**	.69**	.45**	.86**
2. RK: Expertise		–	.81**	.67**	.68**	.65**	.72**	.73**	.55**	.7**	.62**	.83**
3. RK: Genre			–	.72**	.71**	.69**	.74**	.73**	.52**	.74**	.6**	.85**
4. Evidence: Ideas				–	.68**	.77**	.76**	.73**	.49**	.69**	.55**	.87**
5. Evidence: Credibility					–	.7**	.7**	.59**	.45**	.66**	.59**	.78**
6. Evidence: Integration						–	.64**	.64**	.44**	.67**	.74**	.81**
7. Structure: Focus							–	.79**	.58**	.69**	.55**	.83**
8. Structure: Paragraphs								–	.59**	.69**	.57**	.79**
9. Language: Correctness									–	.69**	.48**	.61**
10. Language: Eloquence										–	.69**	.81**
11. Language: Document											–	.66**
12. Holistic score												–

* $p < .05$.
** $p < .01$.

Table 4
Late semester correlations ($n = 53$).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. SAT writing	–	.29	.26	.12	.19	.22	.25	.18	.35	.28	.46**	.29	.32*	.46**
2. RK: Thesis		–	.74**	.67**	.64**	.48**	.59**	.62**	.59**	.53**	.64**	.48**	.84**	.37**
3. RK: Expertise			–	.79**	.74**	.55**	.62**	.67**	.66**	.58**	.71**	.59**	.8**	.24
4. RK: Genre				–	.65**	.55**	.56**	.71**	.71**	.58**	.66**	.61**	.84**	.19
5. Evidence: Ideas					–	.42**	.60**	.7**	.75**	.46**	.54**	.38**	.83**	.11
6. Evidence: Credibility						–	.46**	.53**	.45**	.51**	.54**	.40**	.58**	.37**
7. Evidence: Integration							–	.47**	.46**	.36**	.52**	.71**	.70**	.24
8. Structure: Focus								–	.66**	.57**	.61**	.42**	.78**	.15
9. Structure: Paragraphs									–	.55**	.64**	.37**	.76**	.18
10. Language: Correctness										–	.8**	.47**	.72**	.20
11. Language: Eloquence											–	.66**	.77**	.30*
12. Language: Document												–	.55**	.24
13. Holistic score													–	.23
14. Course grade														–

* $p < .05$.

** $p < .01$.

Table 5
Descriptive statistics and paired sample comparison of early and late semester scores.

Outcome	Group						95% CI for mean difference			
	Early semester (n = 153)			Late semester (n = 153)						
	M	SD	Range	*M*	SD	Range	LL	UL	*t*	*df*
Thesis	5.76	2.54	2, 12	7.44	2.44	2, 12	−2.06,	−1.30	−8.84***	152
Expertise	7.48	2.04	2, 12	8.80	2.17	3, 12	−1.67,	−.974	−7.53***	152
Genre	6.88	2.29	2, 12	8.41	2.12	3, 12	−1.87,	−1.17	−8.53***	151
Ideas	5.96	2.61	2, 12	7.63	2.29	2, 12	−2.06,	−1.28	−8.49***	152
Credibility	6.31	2.55	2, 12	7.93	2.34	2, 12	−2.06,	−1.20	−7.76***	152
Integration	5.57	2.64	2, 12	7.31	2.69	2, 12	−2.16,	−1.31	−8.15***	152
Focus	6.72	2.31	2, 12	8.05	2.27	3, 12	−1.70,	−.714	−6.92***	151
Paragraph	6.97	2.27	2, 12	8.08	2.21	2, 12	−1.46,	−.714	−.5.73***	146
Correctness	8.27	2.32	2, 12	8.66	2.04	2, 12	−.671,	−.113	−.2.78***	152
Eloquence	7.76	2.29	3, 12	8.42	2.20	2, 12	−.947,	−.360	−4.40***	152
Documentation	6.17	3.25	2, 12	8.15	2.85	2, 12	−2.54,	−1.41	−6.89***	151
Holistic	6.07	2.34	2, 12	7.89	2.29	3, 12	−2.16,	−1.47	−10.51***	123

*** $p < .001$.

6. Discussion

This study offers a refreshed concept of validity—in reality, a transformation from validity to a stamp of approval to validation as a process—as a tool to examine the usefulness of consensus statements in local settings. As the evidence suggests, resonance between consensus documents such as the *Outcomes Statement* and their adoption and adaptation at specific institutional sites appears possible. This finding is congruent with that of the University of Cambridge (2011) supporting the use of the Common European Framework of Reference for Languages (CEFR).

For nearly two decades, United States researchers have followed Huot (1996) in his argument that scholars and teachers of writing studies were in the position to "construct a theory of writing assessment based upon our understandings about the nature of language, written communication, and its teaching. . .[and that] these new procedures recognize the importance of context, rhetoric, and other characteristics integral to a specific purpose and institution. The procedures are site-based, practical, and *have been developed and controlled locally* (emphasis added)" (p. 552). His call for localism is congruent with related, more recent calls for attention to results from classroom-based research (Cizek, Rosenberg, & Koons, 2008). In this promising new era of assessment in which the process of validation is undertaken in local settings, Condon (2013) has argued that the "universe of writing assessment" (p. 5) can provide much "richer, more robust assessments," taken as "evaluative occasions that generally do not separate the test from authentic tasks, that provide a rich description of a writer's competencies, and that can allow evaluators to make some judgments about the writer's learning context, as well as about the writer and the writing itself." Because these assessments seek a capacious evaluation of the construct sample, they are positioned to "have the highest *yield* of all" (p. 6). Condon therefore argues that the field of writing studies needs to implement a fuller way to account for validity—one that turns the table on large-scale, standardized, commercial tests. He states "locally developed assessments [can provide] validation studies that meet requisite psychometric standards—evidence based on test content, response processes, internal structure, external variables and consequences. . .[these locally developed], high yield tests should become the standard" (p. 7–8).

But collecting and reporting this information from locally developed programs is easier said than done. For more than twenty years, a uniform system of validation has not emerged in the United States writing studies community. Indeed, in cases where inter-institutional assessments have been attempted, they have done so by concentrating on tasks and rubrics, not on validation processes (Pagano, Bernhardt, Reynolds, Williams, & McCurrie, 2008). The present study suggests that a resolution has been identified to ease the long-standing tension in assessment efforts that try to strike a balance among the requirements of standardization, the demands of educational measurement practices, and the pursuit of localism (Serviss, 2012). In addition, the system we have offered addresses the call by Haswell (2005) for writing studies scholarship that is replicable, data-driven, and able to be aggregated. Along with its potential to resolve tensions and provide empirical evidence, the validation system holds the potential to respond to accreditors and accrediting agencies in the United States as they increasingly require institutions of higher education to document student learning outcomes and provide evidence for how they know their students are learning (Middaugh, 2010; White, Elliot, & Peckham, in press).

While this study provides a promising model that answers Condon's call for locally developed programs to produce "high yield" assessment evidence for validity (2013, p. 107), it is nevertheless important to recognize the substantial complexities involved in validation of score use in local settings. Thus, while Kane's model is a sound foundation, it is also a tool that must, over time, be expanded and modified by the writing studies community in support of evidence-based decision making.

In terms of score interpretation, for example, this study demonstrates the difficulties that arise when trait models are used to score writing samples produced under naturalistic conditions. While adjudicated scores reached the generally accepted .7 correlation range between scorers, non-adjudicated scores did not reach that range. Such findings based on results from this study and others reveal the need for a re-interpretation of the categories to which terms such as "strong," "moderate," and "weak" are attached. While the .7 range may indeed be appropriate for timed, impromptu writing samples, there is no evidence that presently exists that this standard is appropriate for scoring complex writing performances—writing that responds to tasks incorporating construct models such as those in the *Outcomes Statements*—with trait models. That is, there is not a standard set for writing samples written in actual instructional settings—environments in which there are no experimental controls, but the setting in which most students compose their papers.

While the chain of causal evidence runs from scoring to generalization, extrapolation, and implication, the chain of interrogation runs in parallel fashion. Descriptive categories of "strong," "moderate," and "weak" also become questionable when applied to correlations among traits used in support of generalization when the strongest, as Table 3 indicates, does not rise above .85 and correlations between .4 And .6 are frequently observed. Similarly, while regression analysis demonstrates strong relationships among the 11 predictor variables and the outcome variable of the holistic score, for many of the papers in our study, a single rater read each paper for all of the variables. While Table 5 shows a variation in the range of mean scores, a halo effect could nevertheless be in play with the highly related scores an artifact of a the evaluator rather than of the construct itself. Reservations also hold for extrapolation evidence. The SCORE study, as are all present performance studies, taps only the cognitive domain. Yet course grades, as all instructors know, are based on intrapersonal and interpersonal domains (Pellegrino & Hilton, 2012). While the cognitive dimension of writing performance remains of importance, traits such as intellectual openness, work ethic, and conscientiousness (the intrapersonal domain)—as well as collaboration and leadership (the interpersonal domain)—remain of critical significance.

Chiefly of heuristic value, the present study raises key questions about implication, or consequence, the fourth area in Kane's validity framework for interpretation/use argument. We have identified a preliminary list of these questions in Table 6. These questions may be of use as educators align broad consensus statements such as the *Outcomes Statement* and the CEFR with the instructional and assessment needs of specific institutions. As Byrnes (2007) has observed, "Among the many consequences of this age of globalization, migration, multiculturalism, and multilingualism is that governments must address highly complex educational issues. Inasmuch as education is inherently about learners' ability to use language in ways that it is used for schooling, governments are in effect being challenged to devise suitable educational policies" (p. 641). As the present case study demonstrates, Kane's validity framework—to link test scores to claims about their use in an argument-based approach aimed at coherence—appears promising. Whether the efforts are diverse professional initiatives such as the *Outcomes Statement* or internationally sponsored by the Council of Europe, validation frameworks such as the one used in this cases study demand attention to specificity in establishing evidence related to score use.

Table 6
Sample validity questions for local validation of a first-year writing model.

	Scoring	Generalization	Extrapolation	Implication
Basic: essential question	How are inter-reader agreement (consensus) and inter-reader reliability (consistency) established?	How does the writing model map to the curriculum in which it is embedded and assessed?	How does the writing model relate to other measures of writing, from standardized assessments to course grades?	How do results from the assessment impact instructor performance and curricular initiatives?
Operationalization: within the curriculum	Depending on assessment aim, how are acceptable levels of consensus and consistency established?	What methods can be used to demonstrate that the traits of the writing model are related to each other?	What methods can be used to examine the nature of the relationship of the given model to related ones?	In terms of equity and fairness, how does the use of the model impact student diversity in the writing program?
Expansion: beyond the curriculum	What processes are in place to support acceptable levels of consensus and consistency across time?	What methods can be used to demonstrate that the traits of the writing model can predict each other?	What methods can be used to expand the construct model so that its relationship to robust measures may be increased?	How are results from the assessment communicated and used by stakeholders beyond the institution?

These questions in Table 6 have been designed with the premise that a model of the writing construct, such as the *Outcomes Statement*, has been implemented across a given first-year curriculum. (If such a model does not exist, as the present study has made clear, it is impossible to enact validation processes.) While the questions are deliberately limited, they provide a starting place for local validation practices informed by the model presented in this paper. Such programs of research will be essential in exploring and augmenting the role of consensus statements and local context in the age of globalization.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Behm, N., Glau, G., Holdstein, D. H., Roen, D., & White, E. M. (2013). *The WPA Outcomes Statement—A decade later*. Anderson, SC: Parlor Press.

Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett, & W. C. Ward (Eds.), *Construction vs. choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1–27). Hillsdale, NJ: Erlbaum.

Brennan, R. L. (Ed.). (2006). *Educational measurement. Fourth edition. ACE/Praeger series on higher education*. Westport, CT: American Council on Higher Education/Praeger.

Byrnes, H. (2007). Perspectives. *Modern Language Journal*, *91*, 641–645.

Camara, W., & Ernesto, M. (2011, January). Updated standards for educational and psychological testing is released for public comment. *Psychological Science Agenda*, January 2011. Retrieved from American Psychological Association website: http://www.apa.org/science/about/psa/2011/01/testing.aspx

Charlton, C., Charlton, J., Graban, T., Ryan, K., & Stolley, A. F. (2011). *GenAdmin: Theorizing WPA identities in the twenty-first century*. Bloomington, IN: Parlor Press.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, *68*, 397–412.

Collins, R., Elliot, N., Klobucar, A., & Deek, F. P. (2013). Web-based portfolio assessment: Validation of an open source platform. *Journal of Interactive Learning Research*, *24*, 5–32.

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing*, *18*, 100–108.

Condon, W., & Kelly-Riley, D. (2004). Assessing and teaching what we value: The relationship between college-level writing and critical thinking abilities. *Assessing Writing*, *9*, 56–75.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, UK: Cambridge University Press.

Council of Writing Program Administrators. (2000/2008). *WPA Outcomes Statement for first-year composition*. Retrieved from http://wpacouncil.org/positions/outcomes.html

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, *48*, 26–34.

Harrington, S., Malencyzk, R., Peckham, I., Rhodes, K., & Yancey, K. B. (2001). WPA Outcomes Statement for first-year composition. *College English*, *63*, 321–325.

Harrington, S., Rhodes, K., Fischer, R., & Malenczyk, R. (2005). *The outcomes book: Debate and consensus after the WPA Outcomes Statement*. Logan, UT: Utah State University Press.

Haswell, R. H. (2005). NCTE/CCCC's recent war on scholarship. *Written Communication*, *22*, 198–223.

Huot, B. (1996). Toward a new theory of writing assessment. *College Composition and Communication*, *47*, 549–566.

Isaacs, E., & Knight, M. (2013). Assessing the impact of the Outcomes Statement. In N. Behm, G. Glau, D. H. Holdstein, D. Roen, & E. M. White (Eds.), *The WPA Outcomes Statement—A decade later* (pp. 285–303). Anderson, SC: Parlor Press.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Higher Education/Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73.

Little, D. (2007). The Common European Framework of Reference for Languages: Perspectives on the making of supranational language education policy. *Modern Language Journal*, *91*, 646–655.

Maimon, E., Peritz, J., & Yancey, K. (2011). *A writer's resource* (4th ed.). New York: McGraw-Hill Humanities/Social Sciences/Languages.

Malenczyk, R. (2013). Introduction, with some rhetorical terms. In R. Malenczyk (Ed.), *A rhetoric for writing program administrators* (pp. 3–8). Anderson, SC: Parlor Press.

Middaugh, M. F. (2010). *Planning and assessment in higher education: Demonstrating institutional effectiveness*. San Francisco, CA: Jossey-Bass.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing*, *19*, 477–496.

Neff-van Aertselaer, J. (2013). Contextualizing EFL argumentation writing practices within the *Common European Framework* descriptors. *Journal of Second Language Writing*, *22*, 198–209.

O'Neill, P., Adler-Kassner, L., Fleischer, C., & Hall, A. M. (2012). Creating the framework for success in postsecondary writing. *College English*, *74*, 520–533.

O'Neill, P., Moore, C., & Huot, B. (2009). *A guide to college writing assessment*. Logan, UT: Utah State University Press.

Pagano, N., Bernhardt, S. A., Reynolds, D., Williams, M., & McCurrie, M. K. (2008). An inter-institutional model for college writing assessment. *College Composition and Communication*, *60*, 285–320.

Pellegrino, J. W., & Hilton, M. L. (Eds.). (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington, DC: The National Academies Press.

Rhodes, K., Peckham, I., Bergmann, L. S., & Condon, W. (2005). The outcomes project: The insiders' history. In S. Harrington, K. Rhodes, R. O. Fischer, & R. Malencyzk (Eds.), *The outcomes book: Debate and consensus after the WPA Outcomes Statement* (pp. 7–17). Logan, UT: Utah State University Press.

Serviss, T. (2012). A history of New York State Literacy Test Assessment: Historicizing calls to localism in writing assessment. *Assessing Writing*, *17*, 208–227.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9*(4). Retrieved from http://PAREonline.net/getvn.asp?v=9&n=4

Thomas, S. (2013). The WPA Outcomes Statement: The view from Australia. In N. Behm, G. Glau, D. H. Holdstein, D. Roen, & E. M. White (Eds.), *The WPA Outcomes Statement—A decade later* (pp. 165–178). Anderson, SC: Parlor Press.

University of Cambridge. (2011). *Using the CERF: Principles of good practice*. Cambridge, UK: Cambridge ESOL. Retrieved from http://www.cambridgeenglish.org/images/126011-using-cefr-principles-of-good-practice.pdf

White, E. M., Elliot, N., & Peckham, I. *Very like a whale: The assessment of writing programs*. Logan, UT: Utah State University Press (in press).

Williamson, D. D., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practices*, *31*, 2–13.

Wiley, M. (2005). Outcomes are not mandates for standardization. In S. Harrington, K. Rhodes, R. O. Fischer, & R. Malencyzk (Eds.), *The outcomes book: Debate and consensus after the WPA Outcomes Statement* (pp. 24–31). Logan, UT: Utah State University Press.

Yancey, K. B. (2005). Standards, outcomes, and all that jazz. In S. Harrington, K. Rhodes, R. O. Fischer, & R. Malencyzk (Eds.), *The outcomes book: Debate and consensus after the WPA Outcomes Statement* (pp. 18–23). Logan, UT: Utah State University Press.

Diane Kelly-Riley is director of Writing and assistant professor of English at the University of Idaho.

Norbert Elliot is professor of English at New Jersey Institute of Technology,