Measuring the Power of Learning.®

Research Report
ETS RR–17-53

Development and Validation of the Written Communication Assessment of the *HEIghten*® Outcomes Assessment Suite

Joseph A. Rios

Jesse R. Sparks

Mo Zhang

Ou Lydia Liu

December 2017

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Development and Validation of the Written Communication Assessment of the *HEIghten*® Outcomes Assessment Suite

Joseph A. Rios, Jesse R. Sparks, Mo Zhang, & Ou Lydia Liu

Educational Testing Service, Princeton, NJ

Proficiency with written communication (WC) is critical for success in college and careers. As a result, institutions face a growing challenge to accurately evaluate their students' writing skills to obtain data that can support demands of accreditation, accountability, or curricular improvement. Many current standardized measures, however, lack the ability to balance authenticity (i.e., requiring students to produce a sample of writing) with psychometric quality. To this end, we discuss the development of a newly developed measure, the WC assessment of the *HEIghten*® outcomes assessment suite, and present pilot test results based on a sample of 985 test takers from 33 higher education institutions. Overall, we found that the measure includes well-functioning items (i.e., highly discriminating and lacking gender-differential item functioning), an essay task that can be reliably scored by combining human scores with scores provided by an automated algorithm, evidence to support reporting separate selected-response and essay scores to individuals and institutions, and adequate convergent validity evidence. Such results suggest that the HEIghten WC assessment demonstrates promise in providing institutions with a time- and cost-efficient measure of WC that may allow for actionable data to drive decision-making and improve teaching and student learning.

Keywords Written communication; writing assessment; student learning outcomes; validity; automated essay scoring

doi:10.1002/ets2.12185

Proficient written communication (WC) is critical for success in college and careers in the information economy. Individuals must learn how to express ideas, information, and arguments clearly in writing so that these products can be shared with classmates, professors, and eventually, colleagues. A major goal of undergraduate education is to support students in developing proficiency with these skills, making WC a key student learning outcome (SLO) for higher education institutions. This goal is underscored by survey research demonstrating that nearly all chief academic officers consider WC essential for success in and beyond the academy (Association of American Colleges and Universities, 2011; Educational Testing Service [ETS], 2013). In addition, employers value writing skills in the workplace, expressing concerns that college graduates are unprepared for the writing tasks required in the workforce (Casner-Lotto & Barrington, 2006). Accordingly, most employers would like institutions to emphasize development of students' WC skills (Association of American Colleges and Universities, 2011).

Institutions face a growing challenge in the assessment of students' writing skills to obtain information on student learning that can support accountability demands (e.g., accreditation) or internal goals such as curriculum or program improvement, two common uses of assessment information for higher education institutions (Kuh, Jankowski, Ikenberry, & Kinzie, 2014). A survey of provosts illustrates that universities need efficient measures (in terms of time and cost) that can provide actionable data that can inform decision-making and lead to improvements in learning and instruction (Kuh et al., 2014). Their survey revealed that most institutions commonly employ national surveys, rubrics, and locally developed or classroom-based assessments to measure SLOs. Rubrics and classroom-based assessments provide useful formative information to support learning, although questions regarding validity and score reliability may limit the ability to draw strong general conclusions from their data (Jonsson & Svingby, 2007; Reddy & Andrade, 2010). Data from locally developed instruments also cannot easily be compared to data from other institutions. Therefore, an emphasis on assessments that can be scored reliably and administered cost-effectively has resulted in many U.S. institutions' reliance on standardized assessments of writing, which aim to provide both reliable and valid evidence of students' writing skills (Huot, 2002; Murphy & Yancey, 2008). Many such assessments are available; however, these assessments vary in their operational definitions of the writing construct (e.g., Murphy & Yancey, 2008) and in the extent to which they provide valid,

Corresponding author: J. A. Rios, E-mail: jrios@ets.org

actionable information about students' skills (for an extended review of current WC assessments for higher education students, see Sparks, Song, Brantley, & Liu, 2014).

The primary goal of this paper is to provide validity evidence for the newly developed WC assessment of the *HEIghten*® outcomes assessment suite. The initial sections of this paper provide a brief overview of the research base that informed the design of the WC assessment. First, we provide an overview of the WC construct as defined in various higher education frameworks. Second, we briefly review several commonly available writing assessments and their key features. Third, we describe the development of the WC assessment, including an operational definition of higher education-level writing and features of the corresponding assessment design. Although we provide only a brief overview here, the interested reader can find a more detailed discussion of each of these issues in Sparks et al. (2014), which provides the complete conceptual framework for the HEIghten WC assessment. Then, in the second part of the paper, we present results from a recent pilot study designed to provide evidence of reliability and validity of WC test scores, including item quality, automated scoring model validation, dimensionality of selected-response (SR) items, reliability of SR and constructed-response (CR) sections, and WC score relationships with other variables. We conclude with a discussion of the major findings and implications for future research on assessing writing as a SLO.

Definitions of Written Communication

Various organizations have developed frameworks and definitions of WC as a construct to support assessment of students' writing skills. Several notable frameworks providing definitions of WC skills as SLOs include Liberal Education and America's Promise (LEAP) VALUE rubrics (Rhodes, 2010); Degree Qualifications Profile (Adelman, Ewell, Gaston, & Schneider, 2011); European Higher Education Area (EHEA) Competencies (Bologna Framework; European Higher Education Area, 2005); Framework for Learning and Development Outcomes (FALDOS; Council for the Advancement of Standards in Higher Education [CAS], 2009); Framework for Higher Education Qualifications (Quality Assurance Agency, 2008); and the Framework for Success in Postsecondary Writing (Council of Writing Program Administrators [CWPA], National Council of Teachers of English [NCTE], & National Writing Project [NWP], 2011).

Although each framework specifies the WC knowledge and skills students should master by the end of their college career, there is variation in their focus on different aspects of the writing construct (for specific comparisons across frameworks, see Sparks et al., 2014). For example, the Framework for Success in Postsecondary Writing (CWPA, NCTE, & NWP, 2011) is quite comprehensive. The Framework defines five dimensions of writing literacy: rhetorical knowledge (understanding audiences, genres, contexts), critical thinking (analysis, evaluation, synthesis of source materials), writing processes (planning, drafting, editing, revising, responding to feedback), knowledge of language conventions (grammar, discourse organization, tone, style), and the ability to compose in multiple environments (using print and digital modes of production). These dimensions overlap with aspects of writing commonly emphasized across frameworks, including attention to context, audience, and purpose for writing; adherence to conventions of genre (i.e., arguments, expository, narrative); idea organization and content development (i.e., use of relevant and sufficient evidence in supporting an argument); composing in multiple modes and formats (i.e., use of digital composition tools or creation of multimedia products); adherence to language conventions (i.e., grammar, usage, mechanics, and syntax), and style (i.e., word choice, sentence variety, and tone). The considerable overlap among the dimensions suggests some consensus around the value of these core aspects of writing. Further, the dimensions mentioned above are consistent with theoretical models from the cognitive and learning sciences literature (e.g., Bereiter & Scardamalia, 1987; Graham & Perin, 2007; Hayes & Flower, 1980), which view writing as a multifaceted, goal-driven, cognitive, and social process involving the typical writing activities of planning, drafting, and revision, in addition to reasoning about content and audience. Thus, WC requires coordinating social, conceptual, and linguistic representations (e.g., Deane, 2011). Assessments should consider these levels of representation if they aim to provide a more ecologically valid measurement of the writing construct as it is defined both by current cognitive theory and modern educational practice.

Existing Assessments of Writing

Higher education institutions need reliable and valid assessments of WC to obtain a profile of their students' skills. Many existing measures of writing skill are available for use with undergraduates, including assessments designed for course placement, admissions to graduate or professional programs, and SLOs assessments (see Sparks et al., 2014, for an in-depth review of writing assessments spanning these purposes). Here, we briefly review three commonly used SLO assessments that include writing measures: Collegiate Assessment of Academic Progress (CAAP) and ETS Proficiency Profile (EPP), which include both SR and CR measures of writing skill, and CLA+, which measures writing skills using only an extended CR task. Further details on each assessment can be obtained in the aforementioned review and from references cited below.

The CAAP offers two independent paper-based measures—writing skills and writing essay—that institutions may opt to use in combination or separately (CAAP Program Management, 2012). CAAP Writing Skills is a 45-minute test containing 72 SR items, with 12 items associated with each of six reading passages of varied text types (e.g., an editorial); items are embedded in the passage context, assessing students' language conventions (knowledge and skills of grammar, usage, and mechanics) and rhetorical skills (e.g., organization and development of ideas in the passages). On the separate 45-minute CR essay portion, students write argumentative essays on two prompts. Each essay is double-scored by two human raters on a holistic scale (1–6), with the two ratings being averaged to obtain a score for each essay (1–6 in 0.5-point intervals); if raters assign scores that differ by more than one score point, a third rater adjudicates and assigns the final score (CAAP Program Management, 2012, pp. 8–9). Students receive a score for each prompt and a composite score that is averaged across the two essay prompts (1–6 in 0.25-point intervals).

The 72-item SR section yields good school-level reliability (mean of 1,000 random Spearman–Brown adjusted split-half reliabilities = .88; Klein et al., 2009), and ACT reports reliabilities of KR20 = .92 for raw scores (CAAP Program Management, 2012). However, the CR section is less reliable despite the use of two prompts (.75; Klein et al., 2009). CAAP requires 90 minutes of testing time if institutions require both indirect (SR) and direct (CR) writing measures, but it is a fairly robust design. Alternatively, institutions may opt to administer only one section, at a potential cost to validity (because no direct evidence of writing ability is obtained if the CR portion is omitted) or reliability (if the more reliable SR portion is omitted).

The EPP is (in its standard form) a 120-minute suite of assessments which—in addition to sections on reading, mathematics, and critical thinking—includes an SR section of 27 writing skills items assessing students' ability to identify the best revision of a clause or sentence, to organize language to improve coherence and rhetorical effect, and to recognize and revise figurative expressions (ETS, 2010). In contrast to CAAP, which presents SR items in a passage context, EPP writing items are presented within isolated sentences or sentence fragments that must be revised, making them relatively decontextualized (i.e., less valid) compared to items requiring consideration of extended discourse. The EPP SR writing section is moderately correlated with ACT scores ($r = .59$; Banta & Pike, 1989). This section is highly reliable ($\alpha = .91$) and has student-level correlations with the CAAP SR section at $r = .72$ (Klein et al., 2009), suggesting that EPP and CAAP may measure an overlapping construct, despite differences in the contexts in which items are presented. Note that the correlation of EPP with the CAAP essay is much lower, at $r = .33$ (Klein et al., 2009), due in part to limited reliability of CR tasks because of smaller numbers of items within those tasks (Sinharay, Puhan, & Haberman, 2011) and differences in writing construct coverage as measured by direct and indirect approaches (see Murphy & Yancey, 2008). An optional 30-minute essay component assesses students' writing skills with a single prompt scored by an automated engine on a holistic scale (1–6). The prompts present claims about a general-interest topic that can be discussed from various perspectives; students must construct an argument by taking a position on the claim and using reasons and evidence to support it. Correlations of EPP scores with *SAT*® scores range between $r = .27$ and 0.34 across various institution types (Liu, Bridgeman, & Adler, 2012). If institutions choose to administer the optional essay, a total of 150 minutes of testing time is required for the standard form of EPP in addition to the essay. As with CAAP, however, institutions may opt out of the essay section, limiting the construct validity of their results.

Finally, the CLA+, another widely used SLO assessment, includes as one part of the test a 60-minute performance task in which students are presented with a set of several documents and an introductory scenario and asked to write an extended CR in which they analyze, take a position, and possibly draw conclusions about the scenario, using evidence from the source materials as support (Klein, Benjamin, Shavelson, & Bolus, 2007). Essays are scored by at least one human in addition to an automated evaluation system that uses human ratings collected from pilot studies to develop scoring models for each task (Pearson's Intelligent Essay Assessor; Council for Aid to Education, 2015).[1] Students are provided with three subscores (scored analytically, 1–6) for writing effectiveness (i.e., organization and development), writing mechanics (i.e., use of conventions), and analysis and problem-solving (i.e., argumentation and use of evidence). Although this extended CR task addresses many important aspects of the writing construct (including mechanics and effective use of source materials), it lacks reliability when considered apart from SR items ($\alpha = .43$ to .57 across alternate forms; Zahner, 2013), due in part to the use of only one prompt (Sinharay et al., 2011). However, interrater reliability values for the analytic scoring rubric are reasonable ($r = .67$ to .75; Zahner, 2013).

The existing writing measures have several limitations. Construct coverage of SR writing items tends to emphasize relatively low-level skills, such as identifying grammatical, syntactic, or stylistic errors within a sentence (EPP) or passage context (CAAP), rather than prioritizing higher-order skills such as passage-level organization, development, or use of sources. The writing construct is more fully addressed by the CR tasks used in these assessments, which require composing an argument supported by reasons and evidence (although use of sources is addressed only in the CLA+). The writing process itself, however, is not addressed. Although typing an essay requires moment-to-moment revisions of the text with each character produced, which could theoretically be captured and evaluated in online assessments such as CLA+ and EPP, these assessments do not directly capture the revision process or make claims about students' ability to revise or plan their arguments (and CAAP is a paper-based assessment, so only the final written product is obtained). Thus, as with the EPP and CAAP, most writing assessments couple SR and CR components to balance reliability and technical quality (i.e., multiple items providing a more stable estimate of student skills; Sinharay et al., 2011) with authenticity[2] (i.e., a direct writing measure may have greater construct validity than indirect measures; Murphy & Yancey, 2008). However, only the CAAP provides two prompts, whereas EPP and CLA+ each rely on a single CR prompt, making them less reliable (Sinharay et al., 2011). Scoring such CR items also poses a challenge, resulting in a range of approaches. CAAP essays are dual-scored by human raters, increasing the cost and time to score them; in contrast, the EPP's sole use of automated scoring is highly reliable and cost-effective (once a model is built), but limits the inferences that can be made about student writing, given automated scoring engines' emphasis on features associated with *text quality* (i.e., features of the final product) rather than measuring the writer's skill in argumentation, critical use of sources, or other higher-order aspects of writing that are not yet reliably captured using automated approaches (Deane, 2013). By combining discerning human raters who can evaluate high-level writing skills and fast, reliable automated scoring methods that can detect issues with text quality (as is done in the CLA+), construct coverage can be increased while reducing costs due to human raters. Double human scoring may be required only in cases of large discrepancies between the human and the automated scores, where a second rater is required to score only the challenging cases, rather than all essays.

Finally, these existing writing measures require at least 60 minutes to administer if both indirect (SR) and direct (CR) measures of writing skill are desired. Institutions may find it difficult to administer these assessments given this time requirement, particularly if they are to be administered during class time (and if the typical class period is under 60 minutes). This review of the landscape of available assessments (see Sparks et al., 2014, for further details) suggests a need for writing measures that balance authenticity (i.e., CR tasks) with psychometric quality. Such measures should include the kinds of writing tasks that are expected of higher education students (i.e., writing arguments from text sources), that are feasible to administer, and that provide institutions with actionable data on the writing skills relevant to modern definitions of the construct (Deane, 2011). The HEIghten WC assessment module was designed to address these issues using a streamlined design.

Design of the *HEIghten* WC Assessment

The HEIghten WC assessment was designed to measure the WC skills of higher education students as one component of a larger suite of independent SLOs assessments (including modules such as HEIghten Critical Thinking and HEIghten Quantitative Literacy). The WC assessment design is informed by the review and operational definition presented in Sparks et al. (2014). This operational definition of WC, aligned to the frameworks cited previously and research in the cognitive and learning sciences, includes four dimensions: social/ rhetorical knowledge, domain knowledge and conceptual strategies, language use and conventions, and the writing process (see Table 1). In particular, this view of writing emphasizes the coordination of social, conceptual, linguistic, and procedural resources (Deane, 2011) in order to produce high-quality writing at the college level, aligning closely with the LEAP VALUE rubric categories (Rhodes, 2010) of context of and purpose for writing (social/rhetorical), content development (conceptual), genre and disciplinary conventions (social/rhetorical), sources and evidence (conceptual), and control of syntax and mechanics (linguistic). To assess the knowledge and skills as defined in Table 1, the HEIghten WC test has a computer-based design that includes both SR and CR item types (the first three dimensions are measured by SR items, whereas the last is measured, albeit not completely, by the CR task; see Table 2). The 45-minute assessment is designed to be administered within a single-class period while

Table 1 Construct Definition for the Written Communication Assessment

Framework dimensions	Definition
Knowledge of social and rhetorical situations	
Task, context, purpose	The ability to consider and adapt writing to particular purposes (to inform, to argue, to persuade), contexts (academic, professional, social), and tasks
Audience awareness	The ability to effectively consider and adapt one's writing to particular audiences (e.g., experts, nonexperts, specialists, general public)
Genre-specific conventions[a]	The ability to compose texts that adhere to genre conventions (i.e., argument, exposition, essay, critique, summary). For higher education, writing arguments and research reports are common and valued genres.
Composing in multiple modes and forms[a]	The ability to use a variety of technologies (pen and paper, digital software, online environments) to create written products, which may include multimedia elements, particularly when communicating complex information
Domain knowledge and conceptual strategies	
Use of sources and textual evidence	The ability to comprehend and critically analyze a source text (text, data table, image, etc.) and to incorporate information drawn from source texts to develop and support one's ideas, using appropriate attribution
Content development and organization	The development and logical expression of ideas in writing; the ability to fully develop ideas with supporting information and examples from prior knowledge, reading, and experiences; and to present information in a logical, organized, and coherent way
Discipline-specific conventions[a]	The ability to compose texts that adhere to disciplinary conventions specific to one's field of study; related to genre. Includes conventions for source attribution, content, tone, style, organization, and use of evidence.
Knowledge of language use and conventions	
Word choice, tone, voice, and style	The ability to compose text that conveys meaning clearly by using appropriate word choice, sentence variety, tone, voice, and style; what is appropriate will be determined by the context, purpose, and genre
Grammar, usage, syntax, and mechanics	The ability to compose text that is relatively free of errors in grammar, usage, mechanics, syntax, and spelling. Command of the fundamental skills needed to produce fluent text.
Knowledge of the writing process	
Planning, drafting, and revision	Strategic knowledge of the writing process, including prewriting strategies (idea generation, research), drafting, reviewing, revising, editing, and responding to others' feedback. Includes revision in passage contexts.

[a] These aspects are not a focus of measurement in the current written communication test design, which is intended for general use, and therefore examine writing skill in the context of arguments or expositions, digital composition environments, and generic (rather than discipline-specific) writing conventions.

providing both direct and indirect evidence of writing skill. The testing time includes a 20-minute block of SR items and a 25-minute CR section. These sections are described in detail in the following paragraphs.

The 15 SR items are presented in two passage-based sets, which contain items embedded in the context of a passage (under 450 words) and have errors in organization, development, source documentation, linguistic conventions, or other issues (lack of parallel structure, logical comparisons, appropriate use of idioms). Items require test takers to answer questions about the passage (i.e., standard multiple choice items, such as identifying the author's intended purpose or audience for the passage, selecting appropriate references to support claims, or identifying text-structural elements such as relevant supporting evidence) or to revise the text to improve clarity of expression (revision in passage context items, which may ask students to identify the best revision that addresses errors in syntax, grammar, usage, or more substantive organizational issues). These passage-based items include interactive components, such as highlighting (e.g., select the best revision to a highlighted sentence in the passage). Construct-wise, the SR passage-based sets are designed to emphasize the higher-order aspects of writing, with items assessing social/rhetorical knowledge (purpose, audience, and context: 20%) and conceptual knowledge (content development, organization, and use of sources: 40%) comprising the majority (60%) of the SR test items. The remaining 40% of items measure proficiency with lower-level linguistic skills related to word choice, style and tone, or grammar, usage, mechanics, and syntax; these kinds of revision-in-context items can provide evidence

Table 2 Design of the Written Communication Assessment

Section	Length	Item # and format	Scoring	Task types	Subdomains assessed
Selected-response (SR) items (indirect writing measure)	20 minutes	15 MC/SR per form	Automatically scored	Several item types, including revision, use of sources, and development, all presented in passage contexts in varying domains (social sciences, natural sciences, humanities, or everyday/workplace)	Social and rhetorical situations (three items); domain knowledge/conceptual strategies (six items); language use and conventions (six items)
Constructed-response (CR) item (direct writing measure)	25 minutes	One argument essay per form	Essays scored from 1–6 points by one human and e-rater, with additional human ratings if needed	Prompts present a short text describing some issue; students must construct an argument about the issue supported by reasons, evidence, and analysis of the prompt text	Social and rhetorical situations; domain knowledge/conceptual strategies; language use and conventions; writing process (i.e., drafting)

of students' skills in revision (part of the writing process). This design represents an advantage over test designs that focus on sentence-level editing and revision skills (EPP) or that use passage contexts but place more emphasis on lower-level language versus higher-level writing skills (CAAP). The passage-based SR items are embedded in a variety of contexts, including humanities, natural sciences, social sciences, and workplace and everyday situations. In addition, the tone of the arguments presented in these passages includes formal (e.g., academic article) and informal (e.g., trade magazines, editorials) registers. Thus, the HEIghten WC assessment should be feasible to use with general education students, representing a variety of backgrounds and disciplines.

The one-item CR section (which is required) provides direct evidence of composition skill (the drafting phase of the writing process) by asking test takers to compose an argument in response to a prompt featuring a brief argument defending a particular claim attributed to a particular source, such as a letter to the editor of a newspaper. Test takers must compose an original response that adopts and defends a position, using information from the source text and from their own reading and experiences, to develop an organized, logical, and cohesive argument. Test takers must tailor their writing to the specific prompt and purpose (social/rhetorical knowledge); organize and support their argument with additional information and ideas drawn from their background knowledge or from an analysis of the prompt (conceptual knowledge); produce a well-formed text in standard academic English (linguistic knowledge); and plan, draft, and (possibly) revise their argument within 25 minutes (writing process). Scoring uses a combined human and automated approach, with automated scoring leveraging the *e-rater*® V 14.1 automated scoring engine developed at Educational Testing Service (Burstein, Tetreault, & Madnani, 2013), which has been used in a wide range of testing programs for purposes spanning classroom assessment (e.g., *Criterion*®; ETS, 2015) to graduate and professional school admissions (e.g., Ramineni, Trapani, Williamson, Davey, & Bridgeman, 2012). For HEIghten WC, each essay is scored by one human rater and by e-rater; if these ratings are discrepant beyond a 1.5-point threshold, a second human rater will be asked to score the essay, with reported scores reflecting the sum of these human ratings, unless the human ratings differ by more than 1.5 points, in which case a third rater provides an adjudicated score, which is doubled for the final score (the first two ratings are discarded). Further, essays containing unusual response patterns (e.g., empty, non-English, containing excessive grammatical and mechanical errors) will be automatically identified and routed to human evaluation, bypassing e-rater scoring. Taken together, the HEIghten WC design and its streamlined SR and CR sections should provide more comprehensive coverage of the writing construct compared to existing assessments, with measurement of social, conceptual, linguistic, and procedural aspects of writing skill achieved by leveraging both direct and indirect writing measures and by combining human and automated scoring approaches. The rest of this paper describes results of a pilot study designed to provide validity evidence for this design.

Pilot Study to Provide Validity Evidence

Research Questions

The sections that follow describe the evidence collected to support the valid use of the HEIghten WC assessment, based on pilot test data. Specifically, this paper addresses the following research questions:

1. What is the psychometric quality (i.e., item quality, dimensionality, and reliability) of the HEIghten WC assessment?
2. Compared to human – human agreement, how well do automated scores agree with human ratings?
3. In regard to convergent validity evidence, what is the strength of association between test scores from this assessment and those from the SAT Verbal and ACT English assessments for freshman test takers?

Methodology

Sample

The pilot data presented in this study were obtained from administering seven forms of the HEIghten WC module to students at 33 2-year and 4-year institutions ($N = 3{,}464$); however, due to space limitations and the fact that the results between forms were similar,[3] the present study focused on two of the seven forms, as both of these forms were used for validating the automated scoring procedure. In total, the two test forms were administered to 985 test takers, with the majority being recruited from 4-year institutions (89%). At the time of testing, the student level reported by test takers was as follows: 38% freshmen (fewer than 30 semester or 45 quarter hours), 28% sophomores (30 – 60 semester or 45 – 90 trimester hours), 16% juniors (61 – 90 semester or 91 – 145 trimester hours), and 18% seniors (more than 90 semester or 145 quarter hours). In terms of test-taker demographics, students were primarily female (62%), native English speakers (85%), and White (non-Hispanic; 48%); however, 27% were African American, 8% were Latino, 6% were Asian American, and 11% were of other ethnicities (i.e., Native American, Native Hawaiian, or other Pacific Islander) or were multiethnic. Self-reported undergraduate grade point average (GPA) for our sample ranged as follows: 1.00 – 1.99 (1%), 2.00 – 2.49 (8%), 2.50 – 2.99 (23%), 3.00 – 3.49 (35%), 3.50 – 4.00 (30%). Only 3% of test takers did not report an undergraduate GPA, as they were entering their freshmen year; however, for these test takers, their self-reported high school GPA ranged from 2.00 – 4.00.

Measures and Administration

Two computer-based pilot forms of the WC assessment were administered to test takers, who were recruited by their individual institution to participate voluntarily. Test administration was conducted in on-campus computer labs with a trained test administrator providing standardized instructions to all test takers. Each form of the WC assessment comprised 15 SR items and one essay. The 15 SR items were developed to measure three subdomains: (a) knowledge of social and rhetorical situations (three items), (b) domain knowledge and conceptual strategies (six items), and (c) knowledge of language use and conventions (six items). In terms of the essays, the prompts differed across forms. Each prompt presented a short passage in which test takers were asked to construct an argument in response to the claim presented in the prompt. Specifically, Form 1 asked test takers to argue whether the time of day at which school starts should be based on students' internal body clocks (in response to an argument presented to the school board), whereas Form 2 focused on the issue of whether the minimum driving age should be raised to 18 years (in response to a letter to the editor of a local newspaper). This pilot yielded item-level data and total scores. Furthermore, we also collected data to control for students' test-taking motivation, as this was a low-stakes assessment for students, given that the results would have no consequence for them.[4] We collected log file information, such as the number of items not answered and the amount of time spent on each item, and used this information as a proxy for test-taking motivation (Wise, 2009).

In addition to the WC assessment, test takers were administered a computer-based posttest survey. The survey comprised questions regarding self-reported WC proficiency as well as perceptions of test difficulty and time allotted to take the assessment. For the purpose of this study, self-evaluations of WC proficiency were of greatest interest. Additionally, we contacted institutions to obtain both SAT and ACT scores for 652 of the 985 (66%) test takers as well as college GPA for 951 test takers (97%). Although the total testing time allotted for the WC assessment was 45 minutes, test takers were allowed to take as much time as needed on the survey, which took on average 85 seconds ($SD = 57$ seconds) to complete.

Analyses

The analyses conducted in this study were implemented to address three aspects of validation. Firstly, we investigated the psychometric quality of the assessment items by conducting classical test theory (CTT) item analyses as well as test dimensionality and score reliability analyses. Secondly, as this assessment uses both human scoring and automated scoring, we examined the accuracy of the automated scores in relation to human ratings. Lastly, to provide validity evidence for the intended use of scores obtained from the WC assessment, we explored the relationships between WC scores and other related variables, such as SAT, GPA, and self-reported WC skills. However, before addressing these validity concerns, we focused on the issue of low test-taking motivation by identifying and removing test takers deemed to be unmotivated. Each of these analyses is described below.

Motivation Filtering

Two methods were used to identify unmotivated test takers: (a) responding to fewer than 75% of the SR items, and (b) an average response time of 3 seconds or less across all SR items. Data for any test takers meeting one criterion or both criteria were listwise deleted. Upon removing test takers deemed unmotivated by these criteria, validity evidence was collected for the assessment scores.

Item-Level Analyses for SR Items

To evaluate the quality of the SR items on the two forms, CTT approaches to calculating item difficulty, item discrimination (i.e., how well the item differentiates between high- and low-proficiency test takers on the construct of interest), and differential item functioning (DIF; i.e., an item is harder for one subgroup of test takers, for example females, when compared to one or multiple subgroups, for example males, after matching on ability) were applied. Specifically, item difficulty was computed as the mean score (proportion correct; p-value) for the item of interest. Any item that was deemed to be too easy (p-value $\geq .90$) or difficult (p-value $\leq .20$) was flagged as problematic. Item discrimination was computed based on the point-biserial correlation between the item response and the total SR score (r_{pbi}). Any item with an $r_{pbi} < .20$ was flagged as a low-discriminating item. Lastly, gender DIF was examined by applying both the Standardization (STD) and Mantel–Haenszel (MH) procedures (Dorans & Holland, 1993), treating females as the focal group and matching on the total score (combined SR and essay score). Any item with an STD value $\geq .10$ (Dorans & Kulick, 1986) or an MH effect size value (Δ_{MH}) ≥ 1.5 (Dorans & Holland, 1993) was flagged as possessing problematic DIF. Any item deemed to be of poor quality based on the item difficulty, item discrimination, and DIF criteria described above was removed from further analyses.

Dimensionality

Test dimensionality of the SR items was evaluated using confirmatory factor analysis with a three-factor correlated-traits model based on the theoretical structure proposed by substantive writing experts for this assessment (knowledge of social and rhetorical situations, domain knowledge and conceptual strategies, knowledge of language use and conventions). This analysis was conducted in Mplus, version 7.2 (Muthén & Muthén, 2010) using the WLSMV estimator for categorical variables. Adequate model fit was represented by a comparative fit index (CFI) > .90, Tucker–Lewis index (TLI) > .90, and an upper 90% confidence interval value of the root mean square error of approximation (RMSEA) < .06 (Hu & Bentler, 1999).

Reliability

As the intention of the HEIghten assessment is to offer the capability to provide both individual-level and institutional-level scores, we assessed two types of score reliability. Individual-level internal consistency reliability was evaluated based on coefficient alpha for the total (combining SR and CR sections) and subscores. Additionally, institutional-level reliability was estimated using an adaptation of the method proposed by Klein et al. (2009). This procedure involved randomly splitting the students in each school into two samples based on the form they were administered, computing the mean score for each sample within every school, and correlating the institutional means across the two forms for all the schools in the sample. One limitation of this approach was that the reliability estimates may be overly conservative due to the use of half-size samples. To address this limitation, the following Spearman–Brown correction was applied to adjust for the use of half-size samples:

$$r = \frac{2^* r.AB}{1 + r.AB},$$

where $r.\ AB$ is the Pearson correlation between the two samples. As data were only available for 985 students across 33 institutions, any institution that administered a form to 10 test takers or fewer was dropped from the analysis to ensure that the sample size in each institution was not too small. As with the individual-level analysis, both total and subscore reliability were evaluated; however, the school-level reliability approach allowed for the computation of reliability for the essay as well.

Evaluation of Automated Scores

The automated scoring model was developed by employing multiple linear regression of human ratings onto automatically extracted text features such as grammar, usage, mechanics, organization, development, collocation–preposition, word length, sentence variety, and word choice, using e-rater. Of note is that the e-rater features tend to focus on writing fundamentals (i.e., text quality), which are important aspects of writing intended to be evaluated in the HEIghten WC assessment, but do not completely capture full content coverage of the writing construct (e.g., argumentation quality is not measured by e-rater; Deane, 2013). Resulting from the multiple linear regression are (standardized beta) weights assigned to each feature, which serve as the basis of the automated scores.

We evaluated the automated scoring model from three perspectives: (a) agreement between two human raters, (b) the level of balance of e-rater features in contributing to the automated scores, and (c) agreement between e-rater and human ratings. The agreement between human raters was important to evaluate, as the human scores provided an upper limit for the human–machine agreements. According to commonly used industry guidelines, adequate agreement was determined based on quadratic-weighted kappa (QWK) values of .70 or higher, Pearson correlations of .70 or higher, and Cohen's standardized difference (d) values less than or equal to .15 (Williamson, Xi, & Breyer, 2012). The balance level of e-rater features in contributing to the automated scores was evaluated by examining the weights assigned to each feature. Generally, it is undesirable to have a small number of features dominating the automated scoring model. In addition to the same criteria used for human–human agreement, machine–human agreement was evaluated based on the degradation in agreement from human–human to human–machine. As suggested by Williamson et al. (2012), the acceptable levels of degradation were represented as an absolute difference in correlation and QWK between human–human and human–machine ratings of less than or equal to .10.

Convergent Validity

In addition to conducting descriptive analyses and collecting validity evidence based on internal structure, we also evaluated convergent validity by examining the relationship of WC scores with ACT English and SAT Verbal scores. This analysis was conducted only for freshman test takers for two reasons: (a) the difference in time of test administrations was on average shorter when compared to sophomores, juniors, and seniors, and (b) the SAT Verbal (Norris, Oppler, Kuang, Day, & Adams, 2006) and ACT English (ACT, 2009) sections were expected to relate to first-year writing proficiency. As these analyses were not based at the individual item-level as the previously described analyses were, total scores (SR and essay scores combined) were linked across the two forms using the mean–sigma procedure (Kolen & Brennan, 2004). This linking was done for two reasons: (a) the forms were not created to produce equivalent scores (although the forms were parallel in content), and (b) by combining forms, the sample sizes for these analyses were increased, which provided improved statistical power. Upon linking the WC scores, the ACT English scores (1–36) were placed on the same scale as the SAT Verbal scores (300–800) via a concordance table for comparability (ACT, 2015). Combining scores allowed us to increase our statistical power, as data were only available for 124 and 88 test takers on the SAT Verbal and ACT English sections, respectively. Furthermore, previous research has found the relationship between the SAT Verbal and ACT English sections to be strong (r [1,073] = .74, $p < .01$; Koenig, Frey, & Detterman, 2008), which provided some justification for combining scores.[5] Upon linking scores, the combined WC and SAT/ACT scores were then correlated. However, as writing self-efficacy has been found to be an important predictor of writing proficiency (Pajares, 2003), we controlled for self-reported writing competency when correlating WC and SAT/ACT scores. Self-reported WC competency was assessed by administering a 5-point Likert scale, which asked test takers to rate their writing skills from poor to excellent. Upon controlling for self-reported writing proficiency, it was expected that the correlation between scores would be positive and nonnegligible, as all of these assessments measure various aspects of writing. As suggested by Abma, Rovers, and van der Wees (2016), adequate convergent validity evidence is represented by a correlation of around .50 or greater between similar variables.

Results

Motivational Filtering

Based on the low-motivation criteria of not answering 75% of the SR items or averaging 3 seconds across all SR items, only three test takers were classified as unmotivated across both forms. In comparing the mean scores for the total sample (Form 1: $N = 496$, $M = 10.50$, $SD = 3.62$; Form 2: $N = 492$, $M = 12.38$, $SD = 4.22$) and the sample with only motivated test takers (Form 1: $N = 494$, $M = 10.50$, $SD = 3.63$; Form 2: $N = 491$, $M = 12.38$, $SD = 4.22$), no significant differences were found for the two forms (Form 1: $d = 0$; Form 2: $d = 0$). Regardless, all of the following results were based on the filtered sample (i.e., removal of unmotivated test takers).

Item Analyses for SR Items

The results of the item analyses for the SR items are presented in Table 3. The average item difficulty for Forms 1 and 2 was .51 ($SD = .16$) and .62 ($SD = .10$), respectively. At the individual item-level, no items were found to be either too easy ($p \geq .90$) or difficult ($p \leq .20$). In terms of item discrimination, the means for Forms 1 and 2 were .41 ($SD = .10$) and .48 ($SD = .10$), respectively. For Form 2, no item was found to possess an item discrimination lower than .30; however, Items 9 and 10 on Form 1 had item discriminations as low as .24 and .29, respectively. Although these items were not flagged for possessing low discrimination, they were not very strongly related to the remaining items on the same form. As none of the items were found to have significant issues related to item difficulty and discrimination, they were combined into a total score to evaluate DIF. Results of the DIF analyses for both the STD and MH procedures demonstrated that there were no items that favored either females or males unfairly after matching on total score.

Dimensionality

The three-factor correlated-traits model was fit to the sample data separately for each form. Results demonstrated that only Form 2 was found to provide adequate fit (CFI = .93, TLI = .91, RMSEA = .02 [90% CI: 0, .03]); however, upon closer examination, the intersubdomain correlations ranged from $r = .89$ to $r = 97$. Furthermore, the results of the three-factor model for Form 1 were found to be untrustworthy due to the covariance matrix of the latent variables not being positive–definite. The reason for this occurrence was extremely high interfactor correlations (>1.00), which suggested that multicollinearity was present. Due to the very high intersubdomain correlations for both forms, unidimensional models were tested. Results of the unidimensional models demonstrated excellent fit for both Form 1 (CFI = .99, TLI = .99, RMSEA = .02 [90% CI: 0, .03]) and Form 2 (CFI = .99, TLI = .99, RMSEA = .02 [90% CI: 0, .03]), which provided validity evidence for reporting a SR total score. However, Form 1 was found to possess two items that possessed nonsignificant factor loadings.

Although not included in the model, the point–polyserial correlations between the SR total score and the essay were .38 and .48 for Form 1 and Form 2, respectively. The differences in correlations between forms may have largely been due to the problematic items on Form 1; however, the correlation obtained for Form 2 is similar to previous research that has evaluated the relationship between SR and essay sections on English composition assessments (Bridgeman & Morgan, 1996).

Table 3 Results of Item Analyses

	Form 1				Form 2			
Item	p	r_{pbi}	STD (*SE*)	Δ_{MH} (*SE*)	p	r_{pbi}	STD (*SE*)	Δ_{MH} (*SE*)
1	.42	.36	−.07 (.05)	.80 (.49)	.42	.47	−.06 (.03)	−.85 (.37)
2	.76	.51	−.02 (.04)	.51 (.63)	.76	.52	.01 (.03)	.02 (.43)
3	.62	.48	−.05 (.04)	.76 (.53)	.59	.52	.00 (.03)	−.14 (.38)
4	.47	.41	−.04 (.05)	.54 (.49)	.59	.34	.02 (.03)	.20 (.35)
5	.74	.51	.02 (.04)	.24 (.60)	.77	.53	.00 (.02)	.02 (.46)
6	.44	.50	.09 (.04)	.99 (.51)	.64	.38	.02 (.03)	.13 (.36)
7	.34	.45	.01 (.04)	.17 (.53)	.53	.45	.01 (.03)	.06 (.36)
8	.31	.31	.03 (.04)	.11 (.51)	.56	.61	.03 (.03)	.37 (.40)
9	.27	.24	.02 (.04)	.22 (.51)	.68	.60	.03 (.03)	.53 (.43)
10	.32	.29	−.05 (.04)	.49 (.49)	.61	.51	−.01 (.03)	−.01 (.38)
11	.64	.48	−.01 (.04)	.03 (.53)	.48	.58	−.08 (.03)	−1.18 (.40)
12	.56	.41	−.06 (.05)	.73 (.49)	.53	.49	.03 (.03)	.36 (.36)
13	.42	.41	.08 (.04)	.74 (.50)	.66	.45	−.01 (.03)	−.17 (.37)
14	.43	.51	.04 (.04)	.42 (.53)	.66	.51	.02 (.03)	.19 (.39)
15	.69	.56	.05 (.04)	.46 (.58)	.63	.63	−.05 (.03)	−1.01 (.43)

Notes. Items were flagged as being problematic if they were too easy ($p \geq .90$) or difficult ($p \leq .20$), low-discriminating ($r_{pbi} \leq .20$), or possessed gender DIF (STD ≥ .10 and/or $\Delta_{MH} \geq 1.50$). Additionally, the sign of the DIF values was based on using females. Consequently, a positive DIF value indicated that the item was easier for females, whereas a negative value possessed the opposite interpretation.

Reliability

Because the dimensionality analyses provided evidence for a unidimensional representation of the SR scores, reliability analyses were not conducted for the subdomains. At the individual-level, the reliabilities of the SR item scores and total score (combining both SR and CR scores) were found to be adequate for Form 2 (SR scores: $\alpha = .77$; total score: $\alpha = .77$), although these reliabilities were slightly below adequate for Form 1 (SR scores: $\alpha = .66$; total score: $\alpha = .68$). A plausible explanation for the latter result may be related to the nonsignificant factor loadings found for two of the 15 SR items on Form 1. Thus, it is expected that if these items are replaced with better indicators, the reliability of this form will increase. In terms of institutional-level reliability, estimates were based on 19 of 33 institutions, as these institutions administered each form to more than 10 students. Overall, the pilot assessment was found to possess adequate institutional-level reliability estimates for SR (.88), essay (.93), and total (.89) scores.

Automated Scoring Model Evaluation for Essay Items

The automated scoring results are presented below for interhuman agreements, feature weights for the automated scoring model, and human–machine agreements.

Interhuman Agreements

Table 4 provides the agreement statistics between the two human scores and between human and e-rater scores, separately for the two prompts. We noted that the human–human agreements were exceptionally high in this sample based on our a priori criteria. The exact percentage agreements were 71% and 76% for the two prompts, respectively, and the QWK and correlation coefficients were both as high as .88 for both prompts. Further of note is that human raters went through professional training prior to scoring and were calibrated to ensure appropriate and consistent application of a scoring rubric. Hence, we were confident that the human scores could be used as a “standard” to be modeled by the automated scoring system.

Table 4 Model Performance on Cross-Evaluation Sample

		Human1		Human2		Human1 – Human2 agreement					e-rater		Human1 – machine agreement				
Prompt	*N*	Mean	*SD*	Mean	*SD*	*d*	QWK	% exact agree	% adj agree	*r*	Mean	*SD*	*d*	QWK	% exact agree	% adj agree	*r*
1	166	2.98	1.15	2.99	1.16	0.01	0.88	71.08	98.80	0.88	2.99	1.11	0.01	0.86	68.07	98.80	0.89
2	163	2.96	1.01	2.98	1.06	0.01	0.88	76.07	99.39	0.88	2.97	1.06	0.00	0.88	75.46	99.39	0.90

Notes. *d* = Cohen's *d* value; % exact agree = exact percentage agreement; % adj agree = exact plus one-point adjacent percentage agreement; QWK = quadratic-weighted kappa; *r* = Pearson correlation coefficient.

Table 5 Feature Weights in the Final Scoring Model

Feature	Relative weight
Grammar	1.69%
Usage	1.17%
Mechanics	5.46%
Organization	48.74%
Development	32.87%
Collocation – preposition	1.22%
Word length	6.13%
Sentence variety	2.72%
Word choice	0%

Notes. The word choice feature was dropped from the scoring model due to its negative beta weight to the human ratings. The weights shown in this table are based on a model run without the word choice feature.

Feature Weights

The *R*-square of the multiple linear regression scoring model was 0.85. Table 5 gives the feature weights in the multiple linear regression scoring model. We note that the model is dominated by two features: organization and development, which together account for approximately 80% of the reliable variance in the human scores. This finding might be due to the heavy influence of essay length on human scores, because both the organization and development features are closely related to essay length (i.e., number of words). The Pearson correlations between one human score and essay length are .82 for Prompt 1 ($N = 166$) and .83 for Prompt 2 ($N = 163$). One feature—word choice—was further dropped from the final scoring model due to its negative partial correlation with the human scores ($r = -.001$). However, another vocabulary feature, word length, entered into the scoring model and accounted for a notable weight (6.13%).

Human – Machine Agreements

The human – machine agreements are also given in Table 4. For Prompt 1, the human and e-rater score agreement values were 68% exact agreement and 99% adjacent agreement, with a correlation of .89 and a QWK of .86. The comparable values for Prompt 2 were 76% and 99% for exact and adjacent agreements, .90 for Pearson correlation, and .88 for QWK. Cohen's *d* values for human and machine scores were .01 and .00 for the two prompts, respectively. The agreement between e-rater and human scoring also met the degradation criterion of less than .10 from human – human agreements. In fact, the "degradation" shows a marginal increase in terms of correlation: from .88 to .89 in Prompt 1 and from .88 to .90 in Prompt 2. By the standards of e-rater scoring model performance stated previously, both prompts received acceptable e-rater model performance.

Convergent Validity

In addition to evaluating the psychometric quality of the items on the WC assessment and the validity of the automated scoring model, we also investigated the relationship of WC scores with those taken from the SAT Verbal and ACT English sections. In total, SAT/ACT scores were available for 144 freshmen; however, of these individuals, 59 took both the SAT and ACT exams. When this occurred, the assessment with the highest score was used for each test taker. As the ACT scores were put onto approximately equivalent scales (using a concordance table) as the SAT scores, the sample possessed a mean SAT/ACT score of 535.46 ($SD = 111.20$) and a mean WC score of 10.83 ($SD = 4.13$). Upon controlling for self-reported writing proficiency, which was found to be a nonsignificant predictor of WC scores ($\beta = .05$, $p = .48$), the correlation between WC and SAT/ACT scores was r (144) = .47, $p < .001$.[6] Put another way, test takers within this sample with a 1 standard deviation increase in SAT/ACT scores were expected to have a .48 standard deviation increase in WC scores ($p < .01$). The amount of variance in WC scores accounted for by SAT/ACT scores without controlling for self-reported writing proficiency was 23%.

Discussion

The focus of this study was to report on the development and validation of a new SLO writing assessment. To this end, several analyses were conducted that generally showed validity evidence based on both internal structure and relationships to other variables. The following discussion will focus specifically on the adequacy of the items as indicators of the proposed theoretical framework as well as the accuracy of the essay scores and their relationship to the SR section because these results have implications for how scores should be reported to individuals and institutions.

To begin, we found that the assessment demonstrated adequate psychometric quality (i.e., appropriate item difficulty and discrimination between high and low performers, no evidence of gender bias, and reasonable reliability), but the three content domains conceptualized to measure WC were highly correlated, which suggests that students who were high on one domain tended to be high on other domains as well. This finding mirrors similar research conducted by Urbach (2014), in which the scores for the content domains of writing content (i.e., reading engagement, reflection, ideas, and plot), language (i.e., grammar, punctuation, vocabulary use), and spelling were found to possess negligible multidimensionality for data collected from the International Schools' Assessment ($N = 27{,}447$; Grades 3–10). The WC dimensionality analysis result has implications for score reporting and thus points to the need to identify a method that can help detect relative strengths and weaknesses of each domain without providing misleading information when the domains are highly correlated. In addition, we obtained evidence to support reporting separate SR and essay scores at both the individual and institution levels. Such evidence was based on (a) the essay score providing high human–machine score consistency and split-sample reliability estimates, and (b) moderate correlations between the SR total score and the essay. The moderate correlations suggest that the SR items and essay measure similar but not identical skills, which justifies the inclusion of both, and is also supported by previous research (Bridgeman & Morgan, 1996). Furthermore, the scores obtained from the WC assessment were found to possess positive nonnegligible correlations with combined SAT Verbal and ACT English scores ($r = .47$), which may suggest adequate convergent validity evidence.

It should be noted that the findings from this study are limited, as they do not provide direct validity evidence of score-based inferences for the operational WC assessment due to our use of a convenience sample from the pilot test. However, these results will be used to inform and improve the development of the operational assessment. It is important that future data collection efforts for the operational forms allow for conducting further validity analyses, which we were unable to conduct due to the small samples obtained for each form in the current study. As an example, in this study, item-level invariance (or lack thereof; DIF) was evaluated solely for gender. Clearly, for test fairness purposes, it is important to investigate other subgroups, such as underrepresented ethnic minority groups or linguistic groups, which will require larger subgroup sample sizes. In addition, subgroup invariance should also be investigated at the scale-level using structural equation modeling. Such an analysis would add support to investigations of whether institutions can validly compare subgroup means, which may be important for institutional improvement efforts. Additionally, we will need to evaluate whether the automated scoring model exacerbates subgroup differences when compared to human scoring. This is an important analysis, as previous research has found that the e-rater automated scoring engine evaluates certain linguistic and citizenship groups higher than human raters (Bridgeman, Trapani, & Attali, 2012). Further, an alternative modeling approach that can better balance the contribution of different automated scoring features is worth investigation. Therefore, as the operational assessment is developed, these analyses will be need to be taken into consideration. Regardless, this study contributes to the field by reporting information on the development, piloting, and validation of a new writing assessment designed to measure college students' writing ability, in the context of a larger suite of assessments of SLOs. Based on the findings from this study, the HEIghten WC assessment demonstrates promise in providing institutions with a time- and cost-efficient tool that can be used to possibly monitor progress and improve learning of WC.

Acknowledgments

The authors would like to thank Chen Li, Jay Breyer, and Frank Williams from the Educational Testing Service (ETS) for their assistance in preparing the manuscript, as well as Doug Baldwin, Thomas van Essen, Don Powers, James Carlson, David Klieger, and Paul Deane from ETS for their comments on an earlier draft.

Notes

[1] Note that it is not clear which of the two scores is reported to students or how discrepancies in scores between the automated and human ratings are handled.

[2] Note that the design of the CLA+ also includes an SR component to increase the overall test reliability, but these items measure critical thinking rather than writing skills (Zahner, 2013).

[3] The results for all seven forms are available upon request from the corresponding author.

[4] It is expected that the WC assessment will be typically administered in a low-stakes setting for students.

[5] A similar correlation was obtained for 59 test takers in this sample who took both the SAT verbal and ACT English sections, r [57] = .76, $p < .001$.

[6] The individual correlations between WC scores and SAT verbal (r [113] = .44, $p < .001$) and ACT English (r [78] = .44, $p < .001$) were the same when controlling for self-reported writing proficiency.

References

Abma, I. L., Rovers, M., & van der Wees, P. J. (2016). Appraising convergent validity of patient-reported outcome measures in systematic reviews: Constructing hypotheses and interpreting outcomes. *BMC Research Notes, 9*, 1–5.

ACT. (2009). *The ACT writing test technical report*. Iowa City, IA: Author.

ACT. (2015). *International comparative features of the ACT and SAT 2016–2017*. Retrieved from http://www.act.org/solutions/college-career-readiness/compare-act-sat/

Adelman, C., Ewell, P., Gaston, P., & Schneider, C. G. (2011). *The degree qualifications profile*. Indianapolis, IN: Lumina Foundation.

Association of American Colleges and Universities. (2011). *The LEAP vision for learning: Outcomes, practices, impact, and employers' view*. Washington, DC: Author.

Banta, T. W., & Pike, G. R. (1989). Methods for comparing outcomes assessment instruments. *Research in Higher Education, 30*, 455–470.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Lawrence Erlbaum.

Bridgeman, B., & Morgan, R. (1996). Success in college for students with discrepancies between performance on multiple-choice and essay tests. *Journal of Educational Psychology*, *88*, 333–340.

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, *25*, 27–40.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The *e-rater* automated essay scoring system. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current application and new directions* (pp. 55–67). New York, NY: Routledge.

CAAP Program Management. (2012). *ACT CAAP technical handbook 2011–2012*. Iowa City, IA: Author. Retrieved from https://www.act.org/content/dam/act/unsecured/documents/CAAP-TechnicalHandbook.pdf

Casner-Lotto, J., & Barrington, L. (2006). *Are they really ready to work?* Washington, DC: Partnership for 21st Century Skills.

Council for Aid to Education. (2015). *CLA+ technical FAQs*. Retrieved from http://cae.org/images/uploads/pdf/CLA_Plus_Technical_FAQs.pdf

Council for the Advancement of Standards. (2009). CAS learning and development outcomes. In Council for the Advancement of Higher Education (Ed.), *CAS professional standards for higher education* (7th ed.). Washington, DC: Author. Retrieved from http://standards.cas.edu/getpdf.cfm?PDF=D87A29DC-D1D6-D014-83AA8667902C480B

Council of Writing Program Administrators, National Council of Teachers of English, & National Writing Project. (2011). *Framework for success in postsecondary writing*. Retrieved from http://wpacouncil.org/files/framework-for-success-postsecondary-writing.pdf

Deane, P. (2011). *Writing assessment and cognition* (Research Report No. RR-11-14). Princeton, NJ: Educational Testing Service.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*, 7 – 24.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35 – 66). Hillsdale, NJ: Lawrence Erlbaum.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*, 355 – 368.

Educational Testing Service. (2010). *ETS Proficiency Profile user's guide*. Princeton, NJ: Author.

Educational Testing Service. (2013). *Quantitative market research* [PowerPoint slides]. Princeton, NJ: Author.

Educational Testing Service. (2015). Criterion® *online writing evaluation service*. Retrieved from http://www.ets.org/criterion

European Higher Education Area. (2005). *The framework of qualifications for the European Higher Education Area*. Retrieved from http://ecahe.eu/w/images/7/76/A_Framework_for_Qualifications_for_the_European_Higher_Education_Area.pdf

Graham, S., & Perin, D. (2007). *Writing next: Effective strategies to improve writing of adolescents in middle and high schools*. New York, NY: Carnegie Corporation.

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3 – 30). Hillsdale, NJ: Lawrence Erlbaum.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1 – 55.

Huot, B. (2002). *(Re)articulating writing assessment for teaching and learning*. Logan: Utah State University Press.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity, and educational consequences. *Educational Research Review*, *2*, 130 – 144.

Klein, S., Benjamin, R., Shavelson, R., & Bolus, R. (2007). The collegiate learning assessment: Facts and fantasies. *Evaluation Review*, *31*, 415 – 439.

Klein, S., Liu, O. L., Sconing, J. A., Bolus, R. C., Bridgeman, B. E., Kugelmass, H., … Steedle, J. C. (2009). *Test Validity Study (TVS) report*. Washington, DC: Fund the Improvement of Postsecondary Education.

Koenig, K. A., Frey, M. C., & Detterman, D. K. (2008). ACT and general cognitive ability. *Intelligence*, *36*, 153 – 160.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York, NY: Springer.

Kuh, G. D., Jankowski, N., Ikenberry, S. O., & Kinzie, J. (2014). *Knowing what students know and can do: The current state of student learning outcomes assessment in U.S. colleges and universities*. Champaign, IL: National Institute for Learning Outcomes Assessment.

Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher, 41*(9), 352 – 362.

Murphy, S., & Yancey, K. B. (2008). Construct and consequence: Validity in writing assessment. In C. Bazerman (Ed.), *Handbook of research on writing* (pp. 448 – 474). New York, NY: Lawrence Erlbaum.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide: Statistical analysis with latent variables*. Los Angeles, CA: Muthén & Muthén.

Norris, D., Oppler, S., Kuang, D., Day, R., & Adams, K. (2006). *The College Board SAT writing validation study: An assessment of predictive and incremental validity* (College Board Research Report No. 2006-2). New York, NY: The College Board.

Pajares, F. (2003). Self-efficacy beliefs, motivation, and achievement in writing: A review of the literature. *Reading & Writing Quarterly*, *19*, 139 – 158.

Quality Assurance Agency. (2008). *The framework for higher education qualifications in England, Wales and Northern Ireland: August 2008*. Mansfield, England: Author.

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., and Bridgeman, B. (2012). *Evaluation of the* e-rater® *scoring engine for the GRE® issue and argument prompts*. ETS Research Report Series, 2012: i – 106.

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education, 35*, 435 – 448.

Rhodes, T. L. (Ed.). (2010). *Assessing outcomes and improving achievement: Tips and tools for using rubrics*. Washington, DC: Association of American Colleges and Universities.

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice*, *30*(3), 29 – 40.

Sparks, J. R., Song, Y., Brantley, W., & Liu, O. L. (2014). Assessing written communication in higher education: Review and recommendations for next-generation assessment. *ETS Research Report Series*, *2*, 1 – 52.

Urbach, D. (2014). Examining the factor structure of writing assessment based on sets of analytical marking criteria. *Procedia - Social and Behavioral Sciences*, *141*, 1106 – 1111.

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

Wise, S. L. (2009). Strategies for managing the problem of unmotivated examinees in low-stakes testing programs. *The Journal of General Education*, *58*, 152–166.

Zahner, D. (2013). *Reliability and validity of CLA+*. Retrieved from http://cae.org/images/uploads/pdf/Reliability_and_Validity_of_CLA_Plus.pdf

Suggested citation:

Rios, J. A., Sparks, J. R., Zhang, M., & Liu, O. L. (2017). *Development and validation of the written communication assessment of the HEIghten® outcomes assessment suite* (Research Report No. RR-17-53). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12185

Action Editor: Don Powers

Reviewers: Paul Deane and David Klieger