Assessing Integrated Writing Tasks for Academic Purposes: Promises and Perils

Alister Cumming

To cite this article: Alister Cumming (2013) Assessing Integrated Writing Tasks for Academic Purposes: Promises and Perils, Language Assessment Quarterly, 10:1, 1-8, DOI: 10.1080/15434303.2011.622016

To link to this article: https://doi.org/10.1080/15434303.2011.622016

Published online: 25 Feb 2013.
The five studies presented in this special issue offer unique evidence, analyses, and theoretical rationales for assessment tasks that involve writing in reference to information from source material with substantial content. I review the five studies in respect to five “promises” and five “perils,” concluding that, collectively, the promises were mostly fulfilled, but so were most of the perils. The promises are that these task types (a) provide realistic, challenging literacy activities; (b) engage test takers in writing that is responsible to specific content; (c) countertest method or practice effects associated with conventional item types; (d) evaluate language abilities consistent with construction-integration or multiliteracies models of literacy; and (e) offer diagnostic value for instruction or self-assessment. The perils of these task types, however, are that they (a) confound the measurement of writing abilities with abilities to comprehend source materials; (b) muddle assessment and diagnostic information together; (c) involve genres that are ill-defined and so difficult to score; (d) require threshold levels of abilities for competent performance, producing test results that may not compare neatly across different ability levels; and (e) elicit texts in which the language from source materials is hard to distinguish from examinees’ own language production.

When the five studies gathered for this special issue of Language Assessment Quarterly were presented at the Language Testing Research Colloquium on March 19, 2009, in Denver, Colorado, Dorry Kenyon remarked that they move language assessment toward a valuable, new construct definition for academic literacy. I fully agree. These five studies offer unique empirical evidence, exacting analyses, and compelling theoretical rationales for an expanded conceptualization of writing tasks that require learners or test takers to incorporate substantive content from source materials in print, audio, and/or visual forms. These operationalizations conceptualize the assessment of academic literacy abilities in second languages in ways that hold considerable promise.
both practically and theoretically for the evaluation of writing as inherently related to reading and other literate abilities. The studies do not constitute a manifesto, however, for two reasons. First, the researchers have approached these matters independently, and from different perspectives and assessment contexts, highlighting different aspects of integrated writing tasks and their purposes. Second, the authors are careful to observe and analyze the limitations and shortcomings associated with these types of assessment tasks. In doing so, they set a useful agenda for further inquiry.

FIVE PROMISES AND FIVE PERILS

The five studies, collectively, identify five “promises” as well as five “perils” associated with writing tasks that require the incorporation of substantive content from source materials for evaluation purposes. The promises are that these task types

- provide realistic, challenging literacy activities,
- engage test takers in writing that is responsible to specific content,
- counter test method or practice effects associated with conventional item types,
- evaluate language abilities consistent with construction-integration or multiliteracies models of literacy, and
- offer diagnostic value for instruction or self-assessment.

The perils of these task types, however, are that they

- confound the measurement of writing abilities with abilities to comprehend source materials;
- muddle assessment and diagnostic information together;
- involve genres that are ill-defined and so difficult to score;
- require threshold levels of abilities for competent performance, producing results for examinees that may not compare neatly across different ability levels; and
- elicit texts in which the language from source materials is hard to distinguish from examinees’ own language production.

Each of these promises and perils is worth considering for the future design, development, and evaluation of assessments of academic literacy in second or foreign languages. The studies each provide useful data and recommendations to do so.

PROMISES

The primary rationale for writing tasks that require the integration of content from source material is that, fundamentally, this is what writing for academic purposes involves. Students at schools, colleges, or universities are mainly asked to write in order to display their knowledge of ideas and information from reading, listening to lectures, and interacting during courses as well as their abilities to analyze and communicate this material purposefully and coherently relevant to their fields of study (Langer & Applebee, 1987; Leki, Cumming, & Silva, 2008; Rosenfeld, Leung, & Oltman, 2001; Sternglass, 1997). For academic purposes, the construct of writing ability that
needs to be assessed includes the ability not only to compose coherent written texts in conventional formats but also to produce texts that are “content responsible” (Leki & Carson, 1997), that is, related meaningfully and appropriately to ideas and information in academic sources. A longstanding criticism of the brief writing tasks that appear in many conventional, high-stakes tests is that they lack this integral, content-responsible feature of academic communication, and so they encourage students and teachers alike to focus on learning formal or even formulaic aspects of written texts without engaging critically with ideas or information that are central to academic discourse (Hillocks, 2002; Raimes, 1990).

The five studies in this special issue all address this challenge through innovative, exemplary analyses of performance on assessment tasks that integrate writing with reading and/or listening source materials. Each study analyzes how literacy tasks engage students to write realistically in order to interpret, synthesize, and convey substantive information they have read or heard in an assessment context. Sawaki, Quinlan, and Lee (this issue) offer especially sophisticated analyses that reveal three factors (comprehension of source material, productive vocabulary, and sentence conventions) in the integrated writing tasks on the TOEFL iBT that distinguish the writing of learners whose English proficiency is either below or above the level often required for university admissions. Gebril and Plakans (this issue) likewise demonstrate how adult Arabic learners of English vary by proficiency level in their abilities to produce conventional features of written texts in an integrated reading–writing test, observing particularly differences in their abilities to integrate content from source materials. Wolfersberger (this issue), researching a classroom assessment context, shows how students’ task representations are integral to the effectiveness of their writing from sources as well as their teacher’s judgments of that writing. Weigle, Yang, and Montee (this issue) extend these considerations to illuminate the complex, higher order processes of constructing text representations that students use when they must write purposefully about material they have read (rather than respond simply to questions about that material in conventional reading assessments such as multiple-choice items). Yu (this issue) reminds us, further, of the considerable variation that exists in the genre that can be called summary writing, recommending that assessors need to define precisely the intended constructs, skills, task conditions, and scoring methods in integrated writing tasks to produce valid measurements of learners’ abilities (and to avoid confusing variance from factors irrelevant to the construct defined).

These analyses all point the way to new construct definitions of academic literacy for assessment purposes. Most basically, the analyses offer informed alternatives to the dominant convention of assessing language abilities through separate sets of tasks for four so-called skills (of reading, writing, speaking, and listening), instantiated in language testing practices and curricula since Carroll’s (1975) international tests of French, but long recognized to be imprecise and lacking in theoretical justification (Cumming, 1996, 2008). In doing so, the studies build upon, but extend and illustrate for assessment practices, the methods and rationales already proposed for assessing language and content together for students in academic programs (e.g., Byrnes, 2008). Notably, the present analyses also exemplify assessment tasks that may overcome the problem, described by Peirce (1992) among others, of language tests involving static task types that require students merely to display knowledge rather than to communicate authentically or analyze ideas critically. Instead, the present analyses exemplify interactionist theories of language assessment, which assert that test takers’ abilities should be evaluated on the basis of their dynamic performance on “complex assemblies of component information-processing actions that are adapted to
task requirements” and that involve “the intrapersonal process of learning, memory activation, and situated action” (Mislevy & Yin, 2009, p. 250).

As Weigle, Yang, and Montee rightly propose, an obvious theoretical foundation for enhanced construct definitions of academic literacy is Kintsch’s (1998) construction-integration model of text processing, which conceptualizes comprehension as the progressive creation and transformation of cognitive representations at multiple, interacting levels of propositions, texts, and situations. The multimodality of integrated writing tasks extends, moreover, to principles that are consistent with the New London Group’s model of multiliteracies (Cope & Kalantzis, 2000)—viewing literacy as spanning a diverse range of abilities and media including writing and reading, as well as visual and aural representations and new multimedia technologies—a view that has become the basis for most literacy curricula and innovations over the past decade. As Bereiter (2002) has argued, in an age of multimedia literacy and technologies, the focus of educational activities, including assessment, needs especially to be on learners’ processes of knowledge building rather than on the outmoded metaphor of the human mind as a container of objects or repository of traits.

To this end, the crucial, distinctive element in integrated writing tasks is, for test takers and evaluators alike, to focus on uses of written language to construct knowledge effectively from and across relevant sources. This view of writing as a process of purposefully integrating knowledge across communication media also has diagnostic value for instruction and/or self-assessment—in addition to the purposes of assessing English proficiency for university admissions or the completion of course assignments described in the present issue—as suggested by Sawaki et al. (this issue) and by Wolfesberger (this issue). A further benefit for high-stakes testing, observed by Weigle, Yang, and Montee (this issue) and Yu (this issue), is that students cannot really be coached (in a narrow sense) to prepare for integrated writing tasks, or if they are, then they are usefully learning to perform the kinds of writing, reading, and thinking tasks that they will need to do in academic courses.

PERILS

Alas, as I remarked already, despite these promises and their potential, there is no coherent manifesto for the design of integrated writing tasks. There are several constraints on the uses of integrated writing tasks for assessment purposes, which the authors of the present articles all acknowledge but also attempt to understand more fully. The most fundamental constraint is what psychometricians call “task dependencies.” Scores on integrated writing tasks are dependent on test takers’ comprehension of source materials (i.e., abilities to read or listen). Without adequate comprehension of source materials, learners can not write effectively about them, thereby compromising the validity of measurements of writing abilities for learners without adequate comprehension abilities. However, as just observed, this is a problem only if writing is assumed to be an independent “skill,” wholly separate from reading or listening abilities, as Carroll (1975) proposed and as subsequent testing practices have assumed. In contrast, the evidence from research on writing in university and college programs is overwhelming that students’ writing in these contexts does primarily involve their integrating relevant information from source materials that have been read and/or heard. Students’ abilities to do this are what most assessments for academic purposes really do want to evaluate. So the step to take, as the authors all do here, is
to redefine the construct of writing for academic purposes: to include the coherent written communication of interpretations of relevant source material. This step embraces rather than trying to avoid, for psychometric reasons alone, the necessary interdependence of writing performance on reading and/or listening performance. Frankly, is it even possible to write without reading (at least one’s own writing)?

In addressing this issue of task dependencies, the five articles offer compelling reasons to take the radical step toward redefining literate abilities for assessment purposes in academic contexts. Weigle, Yang, and Montee are particularly convincing in demonstrating how students’ reading of source texts deepens, and appears more academically authentic, when they have to interpret the full significance of that source material in order to write about it. Wolfesberger, likewise, describes how crucial students’ appropriate handling of source materials is to their success on written assessments in academic classrooms. From a different perspective, the analyses by Sawaki et al. and by Yu both show how certain language abilities, such as vocabulary knowledge, are interdependent across writing, reading, and listening scores on specific language tests, demonstrating empirically that such abilities span modalities of communication (and may, I suggest, therefore more properly be called a “skill” rather than applying that term to broad performance categories such as writing or reading; cf. Anderson, 1995). Gebril and Plakans (2009, this issue) similarly conclude that students’ abilities to produce certain English discourse features in written compositions relate systematically to their abilities to comprehend and write about source materials, suggesting this ability is integral to their English proficiency for academic purposes.

These latter points lead to the second, related constraint. There are threshold levels of proficiency in a second language above and below which students either may or may not be able to perform integrated writing tasks effectively. The evidence is ample that without sufficient comprehension of source materials second-language writers simply cannot write effectively about them. The implication is that integrated writing tasks are appropriate only for students who have almost achieved or surpassed this threshold level of proficiency in a second language. Sawaki et al. (this issue) suggest that this threshold is precisely the point that tests for admission to college or university programs want to demarcate. To my mind, this is the most powerful implication for assessment policies and practices in this collection, though its validity warrants much further investigation from a variety of perspectives. Admissions decisions to academic programs make sense, in terms of eligibility as well as fairness, only for students who have achieved criterion-level language and literacy abilities to integrate substantial content material into written forms. Accurate assessments of such abilities are also important for other, related purposes such as placement into classes or diagnostic assessments in academically oriented language programs.

The third, fourth, and fifth constraints follow from—and may in fact be evidence for—the notion of a threshold level for effective performance on integrated writing tasks. For this reason these issues warrant consideration together rather than as the separate, itemized points that I presented at the beginning of this article, particularly in view of the recommendations proposed by the authors of the other articles appearing in this special issue. First, integrated writing tasks are difficult to score generically. Yu rightly observes that this may be because the genre of summary writing is itself conceptually amorphous and so often remains ill-defined, is defined operationally in different ways for different language tests, and is inevitably different for tasks with differing source content and writing purposes. Weigle, Yang, and Montee similarly remark on the problem of generating unique rubrics for impressionistic scoring of unique task types, each with unique content, and of establishing interrater reliability for novel tasks. In contrast, Sawaki et al. as
well as Gebril and Plakans make use of established, conventional measures for analyzing written discourse and produce more stable but nonetheless labor-intensive results. Their efforts suggest ways of approaching this scoring problem that may be practical for large-scale assessments; but as Sawaki et al. also imply, doing so in an efficient manner may require automated scoring based on increasingly sophisticated natural language processing research rather than employing human raters of writing.

A further challenge in scoring integrated writing tasks concerns textual borrowing (Shi, 2004). For human raters of compositions, this problem involves first distinguishing, in the written texts produced for such tasks, phrases that derive verbatim from source texts (which may or may not be cited appropriately) from phrases that are the original language of test takers and then, second, judging their appropriateness. This was certainly a challenge in my own research in this area (Cumming et al., 2005), where we found we could not make these identifications reliably and so required a computer program to assume this role mechanically. This problem is remarked upon in all of the articles in the present special issue. More broadly, however, the issue of textual borrowing in academic writing features in nearly every policy and guidebook on composition because it is inevitably a challenge for learning to write in a second language, particularly for those from diverse educational and cultural backgrounds. The pervasiveness of this challenge, and the fine lines involved between literacy learning and academic offence, signal justifications for addressing, rather than avoiding, it in second-language writing assessments. Assessments of writing abilities in higher education really do need to be able to indicate whether students have attained the capacity to use source material appropriately in their writing. But knowing how to make this distinction validly, consistently, and fairly is an area for much further research as well as policy development.

For this reason an important implication from this research is that assessments involving integrated writing tasks may not be informative or useful for administrative decision making in regards to students who have not yet reached the requisite threshold level of being able to perform such tasks effectively. Nonetheless, integrated writing tasks obviously do have pedagogical value for learning these abilities, as Wolfesberger (this issue) and Sawaki et al. (this issue) propose. Writing may, of course, be taught usefully with integrated tasks at any level of language proficiency, so whatever “peril” may be associated with this orientation may reside fundamentally in the nature of high-stakes, formal tests, particularly the convention of establishing linear-type scales to order and report scores sequentially, which is complicated by the existence of a crucial threshold level above and below which examinees’ performance, and so their scores, may differ radically. This may in fact now be a problem for teaching and learning, especially when students are grouped into courses or classes on the basis of common ability levels or learning needs. A final current constraint, however, is knowing best how to approach these matters so as to relate assessments purposefully to teaching and learning. Particularly, how can information from proficiency assessments (what students perform in these contexts) be used for formative, diagnostic purposes (to inform what students need to learn to do next or what instructors should teach these students next)? Investigating these matters is properly the domain of research on assessment in contexts of learning and instruction, rather than formal proficiency testing, as Wolfesberger (this issue) demonstrates. Alderson (2005) and Knoch (2009) have begun to address these matters comprehensively for language tests. Likewise, the research agenda of cognitive-diagnostic assessments has advanced these matters systematically, though more readily for language tests with fixed item-types, such as reading or listening abilities (e.g., Lee & Sawaki, 2009), rather than the more
open-ended, complex, constructed responses that may be interpreted in variable ways as typically appear in the assessment of writing (e.g., Eckes, 2008). Much more inquiry and practical pedagogical advice along these lines remains to look forward to.

In summary, I can observe how the articles in the present special issue point the way forward for future research and development on integrated writing assessments. Most distinctly, this inquiry sets the groundwork for defining new constructs of academic literacy for assessment purposes. In addition to conceptual work on theory building, a fruitful agenda emerges for the refinement of specifications of summary genres and scoring rubrics based on task analyses, further research into second-language learners’ uses of source texts in their written production and development of abilities to do so, identification of threshold levels of text comprehension required for such tasks, validation studies linked to performance in real academic contexts, extensions to research into integrated tasks of speaking abilities on language tests, and analyses of innovative applications of test preparation and diagnostic information in instruction and for learning.

REFERENCES


