The Validity of Examination Essays in Higher Education: Issues and Responses

Article *in* Higher Education Quarterly · July 2010

DOI: 10.1111/j.1468-2273.2010.00460.x

CITATIONS
34

READS
3,451

1 author:

Gavin T. L. Brown
University of Auckland
278 PUBLICATIONS **7,663** CITATIONS

SEE PROFILE

THE VALIDITY OF EXAMINATION ESSAYS IN HIGHER EDUCATION: ISSUES AND RESPONSES

Gavin T L Brown

The Hong Kong Institute of Education

Department of Psychological Studies

Faculty of Education Studies

The Hong Kong Institute of Education

10 Lo Ping Road, Tai Po, NT

Hong Kong SAR

Correspondence

Assoc Prof Gavin T L Brown

Department of Psychological Studies

Hong Kong Institute of Education

10 Lo Ping Road, Tai Po

Hong Kong SAR

Email: gtlbrown@ied.edu.hk

Abstract

The use of timed, essay examinations is a well-established means of evaluating student learning in higher education. The reliability of essay scoring is highly problematic and it appears that essay examination grades are highly dependent on language and organisational components of writing. Computer-assisted scoring of essays makes use of language features and has demonstrated strong similarity to human ratings. Studies of examiner behaviour show that attention to content and language features contributes to grading decisions. However, given the time constraints on essay examinations, an over-emphasis on language aspects may weaken the validity of essay examination grades. This paper suggests alternative approaches to the standard essay prompt that should raise the validity of essay tasks and scoring in higher education. Suggested options include redesigning tasks so that organisational and language features are less influential in scoring and the use of content maps.

Understanding Essay Examinations

Essay examinations demand that students demonstrate knowledge of academic content taught during a course and synthesise an understanding of that content in a relatively brief period of time (Bereiter, 2003; Gentile, 2000). It is assumed that domain experts can and should evaluate the examination essay for its accuracy, completeness, and insightfulness (Bereiter, 2003; Gentile, 2000; Shermis, Koch, Page, Keith, & Harrington, 2002). It has been argued that the evaluation of learning through writing examination essays largely evolved in British-influenced and European education systems (Bereiter, 2003; Murphy & Yancey, 2008).

Examination essays are different to course-work assignments that require the student to produce polished essay papers with support over a period of time from instructors, peers, the library, the internet, and so on. The examination essay also differs from the composition essay commonly used to teach and evaluate writing skills in American colleges. While composition includes content, the emphasis is not so much on demonstrating competence in a content domain, as it is on developing the skills of effective communication and reasoning; in other words, learning to write and writing to learn (Bereiter 2003; Gentile 2000). In contrast to course-related essay examinations, composition focuses on the process of writing, with multiple drafts and much time to write, and on the development of complex cognition (for example, problem solving, decision making, and inferencing) related to communicating with an audience (Hayes, 2000).

Contrary to good writing practice, the examination essay is a first draft piece of writing; it has not been read by a peer, no feedback has been given, and no external tools for editing or proofing were allowed. The essay examination results in a first-draft expression that probably does not represent fairly or accurately the full range of

a student's writing ability or even thinking (Davis, Scriven, & Thomas, 1981). Nonetheless, the essay examination still expects the student to demonstrate a command of a content domain through well-developed written reasoning. The students' written language communication skills are critical in their making clear their knowledge and understanding. However, the conditions of the essay examination suggest that placing a large weight on the quality of written communication may reduce the validity of essay examination scores.

For the purposes of this paper, the focus will be on examination, course-related, essays: that is, timed, on-demand, expositions of knowledge and understanding for summative evaluation. Given the large differences in context and purpose, it seems that processes used to evaluating composition essays or course-work essays may have less validity for essay examinations. Hence, the goal of this paper is to examine the evidence for how higher education examiners read and evaluate examination essays and timed written essay compositions so as to consider whether the validity of essay examinations can be improved. If examiners place too much weight on language features in grading essay examinations, it may well be that essay examination tasks need restructuring to mitigate construct-irrelevant facets in the scoring. This paper assumes that, in higher education, essay scoring depends on professional judges (university professors, lecturers, and tutors) who classify student work according to how well the written essay answers the question prompt posed. It is further assumed that knowledge transforming characteristics (Scardamalia & Bereiter, 1987) and deeper cognitive products (that is, relational and extended abstract processing) (Biggs & Collis, 1982) are associated with highest grades, whereas knowledge telling and surface cognitive products (for example, lists of declarative knowledge) are associated with lower grades. In other words, essay grades ought to reflect not just a quantity of

knowledge accurately reproduced but also the quality of reasoning, logic, or insight the student generates in the time allowed. This quality judgement system is commonplace in academia; for example, doctoral dissertations are judged by panels of domain experts as to whether the candidate has made a substantial and original contribution to the knowledge of a domain and are usually graded as ‘pass as is’; ‘pass with minor changes’; ‘reconsider after major revisions’, or ‘reject’.

Essay examinations are used in higher education because they are intended to generate powerful evidence of higher-order thinking processes and profound understanding of valuable content. Though, it must be kept in mind that the ability of students to produce memorised, pre-prepared essays (consider the Chinese imperial essay examination system that required production of highly prescribed essays [Kaplan, 1972]) and the potential for students to ‘data dump’ everything they know on a topic can undermine the goal of cognitive complexity and depth. The core of the essay is the task, prompt, or question; akin to the stem of a multiple-choice question (Weigle, 2002). The task instructs the student as to the type of writing expected. In most content areas, students will be expected to engage in a certain cognitive task (for example, discuss, compare, contrast, or analyse) in relation to a certain set of content (for example, the causes of World War I, the impact of the setting on a character’s development, or the role of mutation in disease). Such prompts focus attention on important cognitive and content objectives of a course. Given the joint constraints of task and time, it is easy to see that appropriate organisation of material is essential to demonstrate depth and complexity.

Issues in Scoring Examination Essays

Nonetheless, the scoring of content-related essays in higher education has been shown to be notoriously unreliable. The degree of precision, accuracy, and

consistency (that is, reliability—Haertel, 2006) in a set of scores can be estimated three ways: (1) by comparing the degree to which the same essay is given exactly the same or approximately the same scores on separate times or by separate raters (consensus), (2) by finding the degree to which the pattern of high and low scores is similar among markers or markings (consistency), and (3) by establishing the degree to which variance in scores can be attributed to common scoring rather than to error components such as the specific marker or essay task (measurement) (Stemler, 2004). Consensus between essay markers in higher education has been shown to be low. Bell (1980) found exact agreement on a 20-point scale ranged from zero to a maximum of 33%. Scores for the same essay in UK studies varied between 20% and 35% (Caryl, 1999; Meadows & Billington, 2005; Newstead, 2004). The correlation of scores by the same marker, scoring the same essay at different times, has been as low as .37 (Branthwaite, Trueman, & Berrisford, 1981). Reported correlations between different markers have ranged between -.28 and .93 with median values between .40 and .64 (Caryl, 1999; Meadows & Billington, 2005). Unweighted kappa coefficients of only .15 and .21 were found in a study of inter-rater agreement in grading of case study essays in medicine, despite the use of a rubric, suggesting agreement owed much to chance processes (Cannings, Hawthorne, Hood, & Houston, 2005). Student characteristics were only moderately explanatory of variance in medicine essays; suggesting that four markers were needed for each essay to obtain reliable grades (Cannings *et al.*, 2005). Kuper (2006) recommended that with a single marker four to six hours of testing were needed to create dependable essay measurements. As a consequence, Brown (2009) concluded that appropriate rubrics or scoring guides be developed for essay examinations and that, at least, two markers per essay be used. Given the difficulty of reliable scoring, he suggested that alternatives to essay

examinations, including computer-assisted essay scoring, be considered as alternatives.

This line of reasoning raises an important question: what are human raters actually using as the basis for their judgements when they score essays? Barritt, Stock, & Clark (1986) showed that American university composition teachers when faced with first-year student placement essays that followed conventional approaches to writing focused on aspects of the text (organisation, development, use of language, and mechanics). However, when presented with an unusual student essay, such raters also considered who they imagined the writer to be and what the writer might be like as a student of writing. This affective, personal response to the writing may be a function of timing: placement tests take place before instruction and essay examinations, the focus of this paper, occur typically after instruction. Hence, it might be reasonable to expect essay examiners to focus more on the marking scheme and the content. Indeed, Crisp (2008) reported UK examiners responding, albeit infrequently and quite idiosyncratically, with like, dislike, amusement, and frustration to student examination work as just one of their marking behaviours.

More conventionally, Chiang (1999, 2003) found that raters relied heavily on cohesion in judging the overall quality of an essay and specifically found that "transition between sentences in the absence of junction words'' aspect of cohesion provided the best indication of an essay's overall quality rating (Chiang, 2003, p. 480). Eckes (2008), after reviewing a large number of studies of foreign language writing examinations that showed raters do not use marking schemes as intended, found six different profiles in how nine marking criteria were weighted by 65 raters of German as a foreign language. In a study of 300 geography A-level examination essays marked by six examiners, the length of the essay and its focus on the question

were commented on frequently by examiners, no doubt because essay questions leave so much room for digression (Crisp, 2008). Additionally, she found that the same examiners made use of language features (for example, spelling, handwriting, vocabulary, rhetoric) especially when an essay was weak (Crisp, 2010).

Studies of examiner behaviour for various UK GCSE and A-Level examinations (including accounting, business studies, mathematics, physics) have suggested that raters focus mostly on the content or gist of essays when scoring (Weigle, 2002). It would appear markers use a combination of intuitive, automatic approaches (for example, match and scan) and reflective, rule-governed approaches (for example evaluate and scrutinise) while seeking an accurate, appropriately structured, representation of the answer expected by the marking scheme (Sütő & Greatorex, 2008a, 2008b). Further, markers also tacitly refer to what is missing from the expected content even when such absence is not in the official grade criteria (Greatorex, 2000) and heuristics used by markers to simplify complex judgments (for example., knowing the previous marker's grade, the choice of criteria for initial reference) also systematically influence grading (Brooks, 2009). Nonetheless, the more the marker engages in complex, reflective approaches (that is, evaluate and scrutinise), the less accurate the marking (Sütő & Nádas, 2008).

Maclellan (2004) found with 12 Scottish university academics that the assessment of content knowledge was a priority for all and the assessment of higher-order skills of analysis, justification, and evaluation were important to some. However, DeCarlo (2005), from the perspective of signal detection theory, argued that, in the grading of first-year university timed essays, markers lacked the ability to discriminate consistently and accurately between grades and tended to resort to arbitrary and uncontrolled features in deciding the grade to assign. Furthermore, there is reason to

doubt that human raters' essay scores are appropriate measures of knowledge structure (Shavelson, Ruiz-Primo, & Wiley, 2000). Thus, a partial explanation for the low reliability of essay scoring is the differing validity constructs markers use when evaluating an essay.

Thus, there is evidence that while content may predominant in the explicit marking scheme being used, organisation and language are included in the grading decisions of essay examiners. However, it is also clear that the weight given to these factors varies considerably between raters. It is useful to consider research with computer-assisted composition essay examination scoring since it does suggest there are some universal elements in essay scoring which increase the reliability of essay scoring but may limit the validity of essay examination scores.

Automatic Essay Scoring and its Validity for Essay Examinations

Most computer-based essay scoring tools use proprietary and patented statistical algorithms (usually multiple regressions) that use various weighted combinations of linguistic features, vector space of how words are distributed in text, and text structural features to generate a weighted score for an essay (Deane, 2006; McAllister & White, 2006; Shermis *et al*. 2002). Linguistic characteristics used include: (1) syntactic analysis of parts of speech and grammar; (2) vector spaces of key content domain words that appear together having similar form and meaning and distances between them; and (3) text characteristics including such things as: length, number of long words, ratio of passive to active tenses, and organisation. Deane (2006) describes the analytic characteristics of four of the essay rating systems referred to in this paper (that is, project essay grade, intelligent essay assessor, e-rater, intelligent essay grader).

While human scoring is not highly reliable it is still the gold standard against which to evaluate the effectiveness of essay scoring machines (Chung & Baker, 2003). Correlations between the various systems and human judges usually exceed .70 and regularly obtain values in the .80s, while identical scoring on 4 to 6-point holistic scales generally falls in the range 50-65% and approximately equal scoring within one point usually reaches 95−97%. In other words, the machine scoring tools score as consistently as any other pair or group of human raters (Burstein, 2003; Elliott, 2003; Landauer, Laham, & Foltz, 2003; Larkey & Croft, 2003; Page, 2003; Rudner, Garcia, & Welch, 2006). It should be borne in mind that even with these high levels of human scoring emulation, Corso (2006) still found that between 17 and 26% of students were misplaced in first-year college writing classes. Hence, mimicking flawed human scoring accurately will still result in students whose essay grades do not reflect their real ability in writing, let alone their competence in a content domain.

While strong claims of similarity between machine and human scoring are made for tests of writing, independent reports have suggested that some machines are obtaining these high levels of results largely through recourse to the surface features of essays, rather than advanced rhetorical or content dimensions. For example, McGee (2006) found that Intelligent Essay Assessor™ (Landauer, Laham, & Foltz, 2003) scoring was insensitive to sequence of sentences, factual inaccuracy, and jumbled word-orders and mangled syntax. Jones (2006) showed that the IntelliMetric™ (Elliott, 2003) system produced grades primarily based on the length of the essay and its mechanical correctness, rather than on focus, unity, development, elaboration, organisation, structure, or even sentence structure. Page (2003) showed that length is always an important component in scoring with Project Essay Grade. It should be noted that length and mechanical aspects of writing may be legitimate indicators of

content knowledge and cognitive complexity—longer essays have more opportunity to demonstrate deeper understanding of a domain. For example, Jones (2006) suggested that length is a valid indicator of writing fluency, a highly prized attribute in college composition. Nonetheless, the same features may also be a false proxies for correctness, content comprehension, and coherent thought; consider the memorised essay or the 'data dump'.

Improving the Validity of Essay Examinations

It is worth noting that some of the structural or organisational components of essay writing appear to be reasonably readily learned. Marshall (2008), in her study of teaching students to write well-structured introductions to their written essays, found that a two-hour programme was sufficient to teach well-structured introductions to most students in the study. She also found that the score given for the introduction written after the training course predicted a substantial proportion (r=.59; r^2=.35) of the grade given by the student's department for a content-related, course-work essay completed later in the same year.

Thus, because scoring an examination essay may be a response to language skills rather than content or cognitive complexity, there is a confounding explanation for essay scores in examination situations that are meant to evaluate student knowledge and understanding of a content domain. It may be useful to consider whether there are means of reducing the effect of language or surface features on essay examination scores when the goal is coherent, knowledge-transforming explication of deep reasoning. This paper will now suggest some changes to how essay tasks and prompts are written and identify research evidence for focusing on knowledge structure as a valid approach to essay examination.

Reorganisation of Essay Tasks

Since essay scores can reflect organisational features rather than content, it seems useful to mitigate the effect of organisation on the task and the scoring systems. A second reason to do this is that not all students have equal access to the rhetorical organisation patterns used in universities. For example, unless international students adopt a deep-learning approach they might not appreciate the expected discourse structures of academic writing in their international university education (Green, 2007). One approach is to require all essay writers to use the same organisational patterns so that structural characteristics will not be used by the marker as a proxy for knowledge and understanding in the content area. It seems logical to believe that this would generate scores that more closely reflect the intended content knowledge and understanding; however, it has not been possible to locate studies that explicitly test this assertion. Nonetheless, to illustrate the point, two examples are given.

First, in an assignment on understanding the topic of how intelligence has been measured, students in the author's course at the University of Auckland were given the task of analysing the theoretical model underlying a sample set of questions from the XYZ test of intelligence. To guide their response, they were provided a framework for their essay assignment (Task 1).

1. What are the intelligence factors assessed by these items?
   a. Identify the mental ability that each item type is testing.
   b. Explain why the item types group into the factors you have chosen.
   c. What labels from Carroll's taxonomy best describe your factors?
   d. What theorist or theory is most associated with the factor pattern you have chosen?
   e. Why are these factors important measures of intelligence?
2. Explain what kind of relationship you would expect there to be between the factors you have identified.
   a. What kind of correlation and/or factor pattern will there be?
   b. What kind of hierarchy, if any, will there be?
   c. What does this relationship pattern say about the nature of intelligence as measured by the XYZ test?

Task 1. Alternative essay test item (with permission from Brown, Irving, & Keegan, 2008, p. 53)

By following the two main questions and their sub-parts, the students could create a written response that clearly focused on the essential elements of interest, as determined by the content expert, without having to worry that their responses would be scored according to their essay organisation skills. This framework does not control for language selection or mechanical accuracy in the student response, though the importance of the content is foregrounded and the spelling of key content words is modelled. The effect of length is partially controlled in that longer answers focused on each sub-point should exhibit greater knowledge of the domain. More importantly, this framework requires the student to select from all the information they have learned from the course the pertinent information, reducing the tendency to 'data dump'. Further, the framework can be controlled by the examiner preventing students from simply memorising a set response. Just as important, the framework makes use of open-ended questions that should stimulate complex reasoning. All of these factors suggest that, at least on the face of it, this prompt ensures valid responses that can be scored consistently.

A second alternative for course-related essay examinations would be to provide all the sequencing or transition phrases that the student is expected to use. In this way, the content expert as assessor determines the order and content of the essay and the student concentrates on providing relevant content and insightful thought in response to the topic and structure. For example, in an essay, developed by Professor John Hattie, outlining personal views in response to the quotation 'Self-concept and academic achievement are not related' the students were asked to follow the order of a given set of paragraph prompts (Task 2).

> The evidence on this topic generally says …
> Four contrasting findings from the literature on this topic are…
> How do these studies aid in addressing the topic?
> Why is it more beneficial to assess how self-concept relates to learning?
> How does self-concept influence learning, and learning influence self-esteem?
> What strategies do students use to maintain their 'status quo' sense of self-esteem?
> Note some teaching procedures you, as a teacher, could use to redress these strategies.

Task 2. Alternative essay test item (with permission from Brown, Irving, & Keegan, 2008, p. 54)

In this fashion, the students knew exactly what was required of them and they knew exactly what order the material should be presented, both very useful things when performing under time pressure and not having seen the task before. Like the first alternative, this framework mitigates the effect of command of written organisation, de-emphasises mechanical features, and requires students to exercise selection of their knowledge and complex reasoning with that same knowledge. These two alternative approaches to essay examination prompt design have the potential to raise both the reliability of scoring and the validity of essay scores as measures of content knowledge and understanding.

Concept mapping

Since the goal of the course-related essay examination is to identify and evaluate the student learner's knowledge representation of a domain, it seems apt to first develop a clear understanding of the appropriate content for any essay prompt. Presumably, the instructor at the university-level is an expert who can generate an essay response that accurately identifies the expected content and its appropriate relationships for any essay task set. Scoring rubrics and exemplars derived from such responses are likely to contribute to more accurate scoring (Chung & Baker, 2003). Thus, it seems appropriate to insist that before marking begins the course instructor provides the

marking team with an exemplar written answer and a scoring rubric based on the content and knowledge priorities that he or she as the subject expert has. Furthermore, it seems possible that, from the instructor's response, a concept map of the important content and relationships could be constructed.

A concept map (also known as a mind map) is a diagram that displays key content as nodes and key relationships between nodes as annotated paths. Figure 1 is a concept map the author uses to explain his understanding of the purposes of assessment; the map shows that assessment is best defined by its purposes, which can be classified according to their usefulness and the type of use to which they might be put. The key content information (nodes) is displayed in the rectangles, while the annotated directional arrows provide information as to how the various nodes are related to each other. From this map it is possible to reconstruct the gist of the framework for understanding assessment (Brown, 2008). A simple and powerful system for creating concept maps can be obtained from the Institute for Human and Machine Cognition (2010). The point of these maps is that they require knowledge-transformation and at least relational processing to be able to create connections between content learned and understood and the written essay.

Insert Figure 1 about here

The creation of concept maps by American university students before an essay writing examination and its availability during the essay writing reduced construct irrelevant components in student performances (Parkes *et al*. 1999). In a study of teacher education students, the creation of both content maps and written essays helped students improve their mathematics knowledge and learning (Bolte, 1999). Shavelson *et al*. (2005) have argued and shown that assessments that required students to complete concept maps by filling in the node or filling in the paths could be scored

reliably and that they indicated the quality of student learning. An automated essay marking software using concept maps of knowledge content has been developed and the correlation for compare-contrast tasks between the machine and humans, using a generic content rubric, ranged between .45 and .60 (Clariana & Wallace, 2007). While this is low relative to the essay scoring machines that use linguistic features, this may represent the level of agreement possible as long as human raters give so much weight to non-content-related criteria.

Hence, it may be possible to score student understanding directly by evaluation of the concept map they create. Further, it seems likely that requiring students to create concept maps and permitting them to be used in examinations will help their ability to compose written essays under pressure that more accurately reflect their real learning.

Conclusion

The use of timed, course-related essay examinations is a well-established means of evaluating student learning in higher education. While the reliability of essay scoring is highly problematic, what is also of concern is the validity of examination essay grades. It seems evident that, notwithstanding the presence of official marking schemes, much of the marking behaviour of essay examiners is highly idiosyncratic and influenced by many facets of performance not relevant to content learning. It also plain that essay examination grades are highly influenced by the language, organisation, mechanics, and length components of writing. Computer-assisted composition essay scoring makes use of such features and has been shown to be highly able to emulate human ratings of essays. However, other than in the teaching of writing, composition, or rhetoric itself, the main objective of examination essay scoring is probably to determine whether students have a complex, thorough understanding of content.

The conditions under which students write in examinations do not provide a robust basis for allowing scores or marks to be heavily influenced by language skills. This does not mean that full understanding can be shown without good command of written language or writing of a sufficient length. It does, however, suggest that valid marking of knowledge and cognitive complexity in a content domain under examination conditions ought to reduce the effect of communication facets of writing. A key writing skill related to length, content, and cognitive complexity is the explicit organisation of written material. This paper has suggested two alternative approaches to the standard essay prompt that should raise the validity of essay tasks and scoring in higher education. Suggested options included: (1) redesigning tasks so that organisational features are provided to all students and thus mitigate their influence on scoring and (2) the use of content maps as either an adjunct or alternative to essay scoring. While it is clear that these suggestions are not novel insofar as their application in writing courses or in coursework written assignments is concerned, their use in examinations has not been well-established. Higher education needs further studies into validating examination essay grades, one of our primary means of evaluating student learning, and this paper suggests two approaches that may improve the validity of essay examinations.

References

Barritt, L., Stock, P. L., & Clark, F. (1986). Researching practice: Evaluating assessment essays. *College Composition and Communication, 37*(3), 315-327.

Bell, R. C. (1980). Problems in improving the reliability of essay marks. *Assessment & Evaluation in Higher Education, 5*(3), 254-263.

Bereiter, C. (2003). Foreword. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. vii-ix). Mahwah, NJ: LEA.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy (Structure of the Observed Learning Outcome).* New York: Academic Press.

Bolte, L. A. (1999). Using concept maps and interpretive essays for assessment in mathematics. *School Science and Mathematics, 99*(1), 19-25.

Branthwaite, A., Trueman, M., & Berrisford, T. (1981). Unreliability of marking: Further evidence and a possible explanation. *Educational Review, 33*(1), 41-46.

Brooks, V. (2009). Marking as judgment. *Research Papers in Education, iFirst*, http://dx.doi.org/10.1080/02671520903331008.

Brown, G. T. L. (2008). *Conceptions of assessmen*t: *Understanding what assessment means to teachers and students*. New York: Nova Science Publishers.

Brown, G. T. L. (2009). The reliability of essay scores: The necessity of rubrics and moderation. In L. H. Meyer, S. Davidson, H. Anderson, R. Fletcher, P. M. Johnston & M. Rees (Eds.), *Tertiary assessment and higher education student outcomes: Policy, practice and research* (pp. 40-48). Wellington, NZ: Ako Aotearoa.

Brown, G. T. L., Irving, S. E., & Keegan, P. J. (2008). *An introduction to educational assessment, measurement, and evaluation: Improving the quality of teacher-based assessment* (2nd ed). Auckland, NZ: Pearson Education NZ.

Burstein, J. C. (2003). The *E-rater*® scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113-121). Mahwah, NJ: LEA.

Cannings, R., Hawthorne, K., Hood, K., & Houston, H. (2005). Putting double marking to the test: A framework to assess if it worth the trouble. *Medical Education, 39,* 299-308.

Caryl, P. G. (1999). Psychology examiners re-examined: A 5-year perspective. *Studies in Higher Education, 24*(1), 61-74.

Chiang, S. (2003). The importance of cohesive conditions to perceptions of writing quality at the early stages of foreign language learning. *System, 31*, 471-484.

Chiang, S. Y. (1999). Assessing grammatical and textual features in L2 writing samples: The case of French as a Foreign Language. *The Modern Language Journal, 83*(2), 219-232.

Chung, G. K. W. K., & Baker, E. L. (2003). Issues in the reliability and validity of automated scoring of constructed responses. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 23-40). Mahwah, NJ: LEA.

Clariana, R. B., & Wallace, P. (2007). A computer-based approach for deriving and measuring individual and team knowledge structure from essay questions. *Journal of Educational Computing Research, 37*(3), 211-227.

Corso, G. S. (2006). The role of the writing coordinator in a culture of placement by Accuplacer. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 154-165). Logan, UT: Utah State University Press.

Crisp, V. (2008). Exploring the nature of examiner thinking during the process of examination marking. *Cambridge Journal of Education, 38*(2), 247-264.

Crisp, V. (2010). Towards a model of the judgement processes involved in examination marking. *Oxford Review of Education, iFirst*, http://dx.doi.org/10.1080/03054980903454181.

Davis, B. G., Scriven, M., & Thomas, S. (1981). *The evaluation of composition instruction.* Point Reyes, CA: Edgepress.

Deane, P. (2006). Strategies for evidence identification through linguistic assessment of textual responses. In D. M. Williamson, R. J. Mislevy & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 313-371). Mahwah, NJ: LEA.

DeCarlo, L. T. (2005). A model of rater behavior in essay grading based on signal detection theory. *Journal of Educational Measurement, 42*(1), 53-76.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155-185.

Elliot, S. (2003). IntelliMetric™: From here to validity. In Shermis & Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 71-86). Mahwah, NJ: LEA.

Gentile, J. R. (2000). Faculty Forum: An exercise in unreliability. *Teaching of Psychology*, *27*(3), 210-212.

Greatorex, J. (2000, September). *Is the glass half full or half empty? What examiners really think of candidates' achievement.* Paper presented at the annual conference of British Educational Research Association, Cardiff, UK.

Green, W. (2007). Write on or write off? An exploration of Asian international students' approaches to essay writing at an Australian university. *Higher Education Research & Development, 26*(3), 329-344.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65-110). Westport, CT: Praeger.

Hayes, J. R. (2000). A new framework for understanding cognition and affect in writing. In R. Indrisano & J. R. Squire (Eds.), *Perspectives on writing: Research, theory, and practice* (pp. 6-44). Newark, DE: International Reading Association.

Institute for Human and Machine Cognition. (2010, April 11). IHMC CMapTools: Knowledge modeling kit [Web page]. Retrieved from http://cmap.ihmc.us/conceptmap.html

Jones, E. (2006). Accuplacer's essay-scoring technology: When reliability does not equal validity. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 93-113). Logan, UT: Utah State University Press.

Kaplan, R. B. (1972). *The anatomy of rhetoric: Prolegomena to a functional theory of rhetoric*. Philadelphia, PA: The Centre for Curriculum Development, Inc.

Kuper, A. (2006). Literature and medicine: A problem of assessment. *Academic Medicine, 81*(10), S128-S137.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with Intelligent Essay Assessor™. In M. D. Shermis & J. C.

Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87-112). Mahwah, NJ: LEA.

Larkey, L. S., & Croft, W. B. (2003). A text categorization approach to automated essay scoring. In *Automated essay scoring: A cross-disciplinary perspective* (pp. 55-70). Mahwah, NJ: LEA.

Maclellan, E. (2004). Authenticity in assessment tasks: A heuristic exploration of academics' perceptions. *Higher Education Research & Development, 23*(1), 19-33.

Marshall, J. (2008). *Writing introductions in academic essays: The effect of genre-based instruction.* Unpublished dissertation, The University of Auckland, Auckland, NZ.

McAllister, K. S., & White, E. M. (2006). Interested complicities: The dialectic of computer-assisted writing assessment. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 8-27). Logan, UT: Utah State University Press.

McGee, T. (2006). Taking a spin on the Intelligent Essay Assessor. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 79-92). Logan, UT: Utah State University Press.

Meadows, M., & Billington, L. (2005). *A review of the literature on marking reliability* (Commissioned report to the National Assessment Agency). London: AQA.

Murphy, S., & Yancey, K. B. (2008). Construct and consequence: Validity in writing assessment. In C. Bazerman (Ed.), *handbook of research on writing: History, society, schoo, individual, text* (pp. 365-385). New York: LEA.

Newstead, S. (2004). Time to make our mark. *The Psychologist, 17*(1), 20-23.

Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43-54). Mahwah, NJ: LEA.

Parkes, J. T., Suen, H. K., Zimmaro, D. M., & Zappe, S. M. (1999, April). *Structural knowledge as a pre-requisite to valid performance assessment scores.* Paper presented at the annual meeting of the National Council for Measurement in Education, Montreal, QC, Canada.

Rudner, L., Garcia, V., & Welch, C. (2006). An evaluation of the IntelliMetric™ essay scoring system. *The Journal of Technology, Learning, and Assessment, 4*(4), http://escholarship.bc.edu/cgi/viewcontent.cgi?article=1048&context=jtla.

Scardamalia, M., & Bereiter, C. (1987). Knowledge telling and knowledge transforming in written composition. In S. Rosenberg (Ed.), *Advances in applied psycholinguistics, Vol. 2: Reading, writing, and language learning* (pp. 142-175). New York: Cambridge University Press.

Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, E. W. (2005). Windows into the mind. *Higher Education, 49:* , 413-430.

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z., & Harrington, S. (2002). Trait ratings for automated essay grading. *Educational & Psychological Measurement, 62*(5), 5-18.

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

Sütő, W. M. I., & Greatorex, J. (2008a). A quantitative analysis of cognitive strategy usage in the marking of two GCSE examinations. *Assessment in Education: Principles, Policy & Practice, 15*(1), 73-89.

Sütő, W. M. I., & Greatorex, J. (2008b). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal, 34*(2), 213-233.

Sütő, W. M. I., & Nádas, R. (2008). What determines GCSE marking accuracy? An exploration of expertise among maths and physics markers. *Research Papers in Education, 23* (4), 477-497.

Weigle, S. C. (2002). *Assessing writing*. Cambridge, UK: Cambridge University Press.

An earlier version of this paper was presented at the Symposium on Tertiary Assessment and Higher Education Student Outcomes: Policy, Practice, and Research in November, 2008, Wellington, New Zealand. Feedback on this paper from Dr. Paul Hanstedt, Fulbright Scholar at the Hong Kong Institute of Education and Professor of English at Roanoke College, Virginia, is acknowledged.

Figure Captions

Figure 1. Content map of Brown's model of assessment purposes