Measurement: Facilitating the Goal of Literacy

Joanna S. Gorin

Tenaha O’Reilly

John Sabatini

Yi Song

Paul Deane

Educational Testing Service

This manuscript is an early draft of a paper published in Teaching reading and writing and thus may differ slightly from the final published version. Please see below for the official paper:

Gorin, J., O’Reilly, T., Sabatini, J., Song, Y. & Deane, P. (2014). Measurement: Facilitating the Goal of Literacy. In Miller, B., McCardle, P., & Long, R. (Eds), Teaching reading and writing: Improving instruction and student achievement, (pp. 119-128). Baltimore, MD: Paul H. Brookes Publishing Co

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305F100005 to the Educational Testing Service as part of the Reading for Understanding Research (RFU) Initiative. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Abstract

Recent advances in cognitive science and psychometrics have expanded the possibilities for the next generation of literacy assessment as an integrated domain (Bennett, 2011a; Deane, Sabatini, & O'Reilly, 2011; Leighton & Gierl, 2011; Sabatini, Albro, & O'Reilly, 2012). In this paper, we discuss four key areas supporting innovations in assessment for literacy instruction that focuses on reading, writing, and their connection. In particular, we describe how advances in (a) cognitive models, (b) task design, (c) automated scoring and (d) psychometric modeling can work in concert to create a more effective assessment system. First, we argue there is added value in leveraging the relatively separate theoretical research bases on reading, writing, and the emerging literature on their connection to create a unified assessment model for literary. A common model of literacy then enables test designers to develop contextually-rich tasks with items that can be sequenced to help improve not only summative scores, but also provide formative information for students and teachers alike. Coupled with recent advances in automated scoring, current multidimensional and Bayesian modeling techniques appropriate for the complex models can be applied to improve scoring efficiency, accuracy, and instructional utility. The current paper reviews advances and challenges in each of these areas that must be considered in concert for proper design of literacy assessment tools.

Measurement: Facilitating the Goal of Literacy

Assessment, when properly designed, scored, and interpreted, has the potential to play a key role in the educational process (Bennett, 2011a; National Research Council, 2001). Summative assessment is widely-used to make inferences about individual and groups' proficiency at various points in the educational process. Broad initiatives, including Race to the Top (U.S. Department of Education, 2009), have emphasized the need for valid and reliable measures that provide accountability for what students have learned and can do at the end of each K-12 academic year. Much of these efforts have focused on research and development efforts focusing on summative assessment – status or growth assessment that documents the outcomes of instructional experiences. To a lesser extent, efforts have been made to improve the quality and utility of formative assessment tools – measurement, instructional, and professional development tools that use test scores to help inform the instructional process, thereby enhancing student learning. Whether developing formative or summative educational assessments, the success of the assessment in facilitating learning is based entirely on the alignment between the goals of the assessment (i.e., the intended use of the test) and its design and scoring (Gorin, 2006, 2012; Mislevy, 2004). All assessment development activities should proceed from a comprehensive model of the targeted skills that incorporates instructional practices and developmental theories about proficiency and expertise in the domain.

As evidence mounts to support the need for learning and instruction and the reading-writing connection in an integrated framework of literacy, our assessments must be reconceptualized and redesigned such that they are sensitive to this new purpose. Graham and Hebert (2010, 2011) conducted meta-analyses of empirical studies showing that writing about materials one reads can improve students' reading comprehension, reading fluency and word reading. There are also theoretical and practical reasons, and some evidence (e.g., Graham, 2000; Krashen, 1989), to suggest that reading instruction can also enhance writing skills. On the theoretical side, there is some scholarly exploration of social and cognitive models of reading and writing development and integration (e.g., Bazerman, 2004; Bereiter & Scardamalia, 1987; Hayes, 1996; Langer, 2001; Olson, 1991; Olson & Hildyard, 1994; Shanahan, 2006). On the practical side, there are many instances of integration within classroom practices – teachers assign students to write about what they read; students write documents modeled after examples they have read; peers critique each other's writing, and so forth. Despite this growing evidence-base and the common sense observation that reading and writing are inextricably intertwined in one's language and literacy skills, reading and writing are often taught as separate subjects, and this practice is likely reinforced by requiring students to take separate high stakes assessments. If we want an assessment to support instruction and research on students' reading and writing that not only accounts for but leverages their connections, we must build an assessment system with:

- a theoretical model of reading, writing, and their connections,
- tasks that provide behavioral evidence for claims about both reading and writing,
- automated scoring approaches that allow for efficient, valid, and reliable reporting for formative and summative purposes, and
- complex psychometric models that account for the multidimensionality and dependencies that exist among reading and writing

An Integrated Literacy Construct Model

The driving engine for any educational assessment is the definition of the construct in which the desired inferences and claims about student abilities are laid out (Mislevy, 1994).

Recently, the use of cognitive models has been advocated as a powerful tool for construct definition in the assessment design process (Gorin, 2006, 2007; Gorin & Embretson, 2012; Leighton & Gierl, 2011). When moving towards assessment of the reading-writing connection, it is critical that we begin with an appropriate cognitive model of the intended construct – that is, a model of the reading-writing connection itself. Assessments designed to measure reading, writing, and their relationship must therefore begin with an integrated model of the two constructs that includes the nature of the relationships between them and their development.

Deane, Sabatini, & O'Reilly (2011) have drawn on the reading-writing connections literature to develop an ELA competency model, a form of cognitive model, for assessment design that puts a strong emphasis on literacy practices as integrated, socially situated activity systems that should be assessed within the complex array of expressive, interpretive, and deliberative/reflective skills, which are hypothesized to call upon shared, integrated mental representations.[i] Consider any relatively complex, but commonplace literacy activity a student might be called upon to perform - prepare for a class discussion; study for an exam; write a research report or argument; give a presentation on a topic or issue – and one will quickly observe a complex mixture of reading, writing, and thinking skills must be deployed in each. Texts must be sourced, read, and comprehended. Not only must a full-fledged composition be iteratively drafted, but during the writing process ancillary writing skills are deployed for such purposes as notetaking, glossing texts, creating lists, writing summaries or outlines, communicating with peers or querying web sources, and so forth. Throughout this process, reflection, deliberation, and discussion are deployed to reason, articulate, and communicate about ideas initially read, written, or thought. Thus, one of the advantages and contributions of a combined ELA model is that it reinforces the common, shared cognitive resources deployed in literacy activity systems, whether the channel/modality itself is primarily reading, writing, or reasoning. Drawing on this framework, one can design assessments that systematically probe reading, writing, and thinking, while providing insights into how these processes are related.

Let us consider a particular ELA skill that is often assessed separately in either reading or writing - argumentation. In reading, argumentation is generally defined as the ability to identify people's positions, arguments, and evidence; and in writing, as the ability to express one's own position, argument, and evidence in writing (Graff, 2003; Hillocks, 2010, 2011; Kuhn, 2005; Newell, Beach, Smith, VanDerheide, Kuhn, & Andriessen, 2011). The argument component of Deane et al's (2011) integrated ELA competency model targets key argumentation skills in four critical aspects (i.e., appeal building, taking a position, providing reasons and evidence, and framing a case) at five hypothesized developmental levels across the pre-K through college student continuum. It not only specifies how the reading, writing and critical thinking skills related to argumentation shift qualitatively as students achieve higher levels of sophistication, but also identifies strategies that teachers could use to help scaffold students toward the next level – across both reading *and* writing. As skillful writing presupposes a baseline level of reading skill, the integrated argumentation model explicitly includes critical reading skills. In effect, the critical reading tasks specified in the design define receptive prerequisites to success completion of expressive writing tasks. More generally, critical evaluation of content is fundamental to argumentation, and is necessarily involved in planning, writing, and revising processes. Effective writing requires that the writer thinks effectively and critically about the goal, audience, position, and argument. In essence, the integrated argumentation model captures the insight that successful writing is closely integrated with general literacy and thinking skills.

Scenario-Based Task Design

Once a researcher or test developer specifies a construct model, the next crucial issue is task design. What item types or task types should appear in an assessment of ELA literacy? When considering item types for any assessment, one must take a step back and consider a more fundamental question – namely, the behavior types that constitute the strongest evidence of the literacy skills to be measured? One must then create opportunities for such evidence to be observed via test questions and tasks. Of utmost concern are the two most commonly cited threats to validity – construct underrepresentation and construct irrelevant variance. Either of these factors threatens to reduce not only the quality and accuracy of our score interpretations for summative purposes, but also the appropriateness of any instructional decisions made on the basis of formative scores and data. If separate measures of reading and writing are used as a basis for decisions about an integrated literacy curriculum, or to make summative conclusions, the most obvious threat to validity is underrepresentation of the more complex integrated literacy construct. By measuring reading and writing in an integrated assessment, these and other threats to validity can potentially be mitigated and the instructional implications of the assessment can be enhanced. Below we describe some of these potential threats and propose several design features that can be used to address these concerns.

Probably the biggest threat to validity for reading and writing assessments concerns the student’s level of background knowledge on the topic of the texts and writing prompts. In short, students who know more about the topic of the passages and prompts understand and can write more than students who know very little about the topic (Benton, Corkill, Sharp, Downey, & Khramtsova, 1995; O’Reilly & McNamara, 2007; Shapiro, 2004). In the context of assessment, background knowledge poses a serious threat for the interpretation of reading and writing scores. When students have high knowledge, it is very difficult to determine whether the test scores reflect true reading and writing ability or the level of the students' background knowledge on the topic of the text and prompts.

One approach to reduce this threat to validity is to explicitly measure and support the development of background knowledge. Rather than asking students to write from generic prompts, we propose providing students with a collection of texts from which they can write. For instance, students might be asked to write an argument for or against manufacturing hybrid cars. However, prior to writing, students are provided with a set of texts and materials that describe how hybrid cars work, the advantages and disadvantages of producing and driving them on the road and the long term environmental impacts. The texts and readings become the common "background knowledge" that allows students to write on a more level playing field. By asking comprehension questions about the texts, we can also get an estimate of whether students understood the content and whether the lack of understanding presented any barriers for writing a quality essay.

A second threat to validity concerns the potential for narrowing the construct of reading and writing due to lack of a specific purpose for the assessment task. Reading and writing are purpose-driven activities that should represent the range of contexts in which students read and write (McCrudden, Magliano, & Schraw, 2011). However, in a typical reading assessment, the only purpose for reading is to answer multiple choice questions correctly (Rupp, Ferne, & Choi, 2006). Clearly, people read and write in both academic and non-academic settings for a wide range of purposes. Sampling from a range of these purposes in a joint reading and writing assessment not only improves construct coverage and the potential authenticity of the assessment, but it also helps clarify task demands.

To address this second potential threat to validity, we advocate measuring reading and writing skills jointly under the context of a scenario (Bennett, 2011b; Bennett & Gitomer, 2009). A scenario includes an authentic purpose for reading about a collection of sources on a particular topic that culminates in an integrative writing task. For instance, students could be given multiple texts about e-waste, its sources, and opposing arguments about its consequences. The student might be asked questions about the content of the texts and then asked to write to an electronics company about the potential dangers of e-waste and to suggest some steps the company can take to reduce the amount of e-waste. By introducing a critical reading task, we not only measure students' ability to understand text, but also give students content to consider (e.g., to summarize, to analyze, to synthesize, to evaluate) in preparation for writing, thus modeling the kinds of activity systems and processes we aim for students to learn. Further, this design not only builds up students' knowledge of the topic through reading, it also isolates what parts they did or did not understand which is useful for evaluating the quality of their essay. A more sequenced sets of tasks as described here might shed light on areas in which a particular student or group of students may need further instructional support.

Human and Automated Scoring

One of the most significant challenges to any assessment that includes a writing component is that of scoring. Issues of reliability, validity, cost, and time are among the potential limitations of human scoring. No matter how well defined the construct, or well constructed the task, if the scoring cannot be reliable, valid, cost-feasible, and sufficiently rapid, the instructional utility of the assessment scores is limited. Despite these challenges, most (if not all) state writing and reading assessments now include various types of constructed response items. This condition has been facilitated by technological advances that support automated scoring of constructed responses, including automated essay scoring systems such as *e-rater*® and *c-rater*™ (Shermis & Burstein, 2003). These engines have been used to generate scores for summative assessment, as well as feedback for students on their writing quality in formative assessment systems, like the *Criterion*® online writing evaluation system. A growing literature documents the reliability and validity of human and automated scoring of performance items (e.g., short and extended constructed response item) on either reading or writing assessment (See Shermis & Burstein, 2013). We briefly consider the issues of human and automated scoring of constructed responses to integrated literacy assessment.

Human Ratings of Writing Quality

The relationship between reading skill and the ability to produce high-quality writing is complicated by several issues related to scoring, namely, the context-sensitivity, cognitive encapsulation, and variety of the factors that underlie judgments of writing quality.

Context sensitivity. As a construct, writing quality is mediated by the intended audience, and hence by factors relevant to reading. A poem is not read the same way as an essay; nonfiction is not read the same way as fiction; technical manuals are not read the same way as newspapers; thus, the judgments people make about the quality of a piece of writing are necessarily relative to assumptions about its purpose, audience, and genre. These considerations provide reasons to expect linkage between the factors that contribute to text readability and those that indicate writing skill, but not in any simple way. A functional connection between reading and writing guarantees that there will be a large common substrate of abilities shared between reading and writing (Shanahan, 2006), but implies that they will not be deployed in the same way; in fact, features that predict that a text will present greater reading difficulty often also predict higher levels of writing skill (McNamara, Crossley, & McCarthy, 2010).

Cognitive encapsulation. As a performance, skilled writing must take many different factors into account. But the more skilled a student becomes, the more likely that important subskills of reading and writing will be fluent, e.g., both automatized and accurate (Logan, 1997; McCutchen, 2006; Kuhn, Schwanenflugel, Meisinger, Levy, & Rasinski, 2010). A rater is being asked to evaluate how well a reader or writer has controlled a variety of elements that are not normally available for conscious inspection. Such analysis exposes the rater to all the dangers of introspective analysis which leads to low levels of inter-rater reliability, and difficulty separating out traits on writing rubrics, (Elliott, 2005).

Variety of Underlying Factors. Judgments of writing quality are sensitive to a variety of properties of the text, ranging from easily-measurable features such as spelling errors, up to much more global, inferred features such as rhetorical effectiveness and validity of arguments. An early study of rater behavior, French (1962), identified six such factors; descendants of this kind of analysis inform scoring methods such as the 6-trait model (Spandel, 2004) to this day, identifying such writing traits as mechanics/conventions, word choice, organization, and content. However, many of these factors appear to be directly linked to shared skills that may be relevant to reading. For instance, mechanics-related writing abilities are closely linked to mastery of the orthographic patterns needed for effective reading; word-choice-related writing abilities are closely related to vocabulary skills needed for effective reading comprehension, and so forth.

Automated Scoring Methods

Many but not all of the dimensions of writing quality are amenable to automated measurement through the use of automated scoring technologies, which in addition to the speed and cost-efficiency they offer, may resolve some of the above listed challenges to human ratings. The kinds of analysis possible can be summarized by considering the kinds of features employed in a typical automated scoring engine. One such, the as *e-rater*® scoring engine, is well-documented (Attali & Burstein, 2006; Burstein, Chodorow & Leacock, 2003), and can be used to stand in for the larger class. Quinlan, Higgins and Wolf (2009) mapped individual as *e-rater*® features onto the dimensions of a 6-trait writing quality construct. Attali and Powers (2009) demonstrate a factor structure for these features in which many of the features map onto factors that roughly correspond to vocabulary (word choice), accuracy (mechanics/conventions), and fluency (organization and development). A regression model is built using these features to predict human-assigned scores. In general, such models provide strong prediction of human ratings (Burstein & Chodorow, 2010; Chodorow & Burstein, 2004; Powers, Burstein, Chodorow, Fowles, & Kukich, 2001).

Since automated scoring methods are trained using the scores assigned by human ratings, many of the issues of interpretation that arise with human ratings persist when we consider automated scoring models. However, to the extent that the underlying features capture dimensions of writing performance that reflect shared components with reading, it is possible that automated writing analysis methods can contribute to developing a clearer understanding of the relationship between reading and writing. Further these scores can be used to provide meaningful feedback to students about their writing and teachers about their instructional effectiveness. If assessment is to become an integral part of literacy instruction, research on the use of these automated systems to provide timely and instructionally relevant feedback for teachers, not to mention reliable and valid summative scores, is still needed.

Advanced Psychometric Models

Ultimately, in order to make inferences about our claims using observed evidence, the data must be translated into interpretable form. As the constructs we measure and claims we want to make from our assessments increase in complexity, as is the case with an integrated literacy assessment - our analytic tools must also adapt (Gorin & Svetina, 2012; Gorin & Svetina, 2011; Rupp, 2012; Wilson & Moore, 2012). Psychometric approaches, ranging from classical true score theory to item response theory, offer a variety of methods for converting individuals' behaviors into estimated ability levels. In the traditional assessment paradigm the focus is on transforming scored item responses into latent trait estimates, traditionally on unidimensional latent trait scales. Unidimensional models that assume a single underlying latent trait affecting task performance are overly simple for highly contextualized tasks appropriate to measure literacy. With the increased computing power of the last several decades, we now have multidimensional alternatives that are likely exactly what is needed for assessing the reading-writing connection as we have described thus far. We briefly review multi-dimensional modeling approaches that offer the most promise for integrated literacy assessment: multidimensional item response theory and Bayesian inference models (BINs).

Multidimensional IRT (MIRT) allows for the contribution of two or more constructs to the solution for an item or set of items. MIRT decomposes the unidimensional person parameter into an item-dependent linear combination of latent traits (Junker & Sijtsma, 2001, pg. 259). One form of MIRT models, compensatory MIRT allows for an examinee to correctly respond to an item when some, but not all of the skills needed to answer the question are mastered. That is, a high level of ability on one skill may compensate for lower level ability on a different skill, both of which are associated with that same item. However, according to Junker and Sijtsma (2001), while these models made advances in blending IRT and cognitive assessment, they are not sensitive to all aspects of cognition. Thus, noncompensatory MIRT approaches might be more appropriate. In these models, performance on tasks involved the conjunction of successful performances on multiple subtasks, where each subtask may be thought of as unidimensional IRT model.

Though MIRT models allow simultaneous consideration of multiple abilities, these models still typically only model a single piece of observed data - an item-level score. If we follow the advice given to this point in the chapter, the complex tasks for integrated literacy assessment would yield multiple scored behaviors, each of which might be related to one or more of the skills of interest. As the number of observable variables (scores) for an item increases the use of the traditional IRT and MIRT models that typically handle one variable per item are insufficient. One approach that has had some success is the use of Bayesian inference networks (BINs; Jensen, 1996, 2001; Pearl, 1988). BINs are graphical models in which we transmit complex observational evidence within a network of interrelated variables – the skills of interest (unobservable variables) and the scores from the complex task (the observed variables). Conditional relationships between the observables variables, the unobservable variables, and characteristics of the task are graphically diagramed as a network. Then, using Bayes theorem of conditional probabilities, the strength of the relationships and the fit of the overall model to the data can be tested. The key, of course, is to construct a BIN based on a non-arbitrary model of cognition and the tasks – hence the need for a strong theoretical model of the constructs and appropriately designed tasks that are aligned with one another.

Conclusion

As the role of assessment in education increases, it has the potential to guide instruction – for good or ill. For that reason, it is critical that our assessments be designed to measure constructs as we believe that they exist, develop, and are learned. Reading-writing connection requires a complex assessment with the explicit purpose of measuring both constructs within a single system. The development and relationship between students’ reading and writing abilities should be measured by design. It is only in doing so that we can provide educators and researchers with the assessment tools necessary for them to improve reading and writing instruction and learning, an outcome that benefits us all.

Construct models, task design, scoring, and psychometric modeling – advances in any one of these areas – is only effective if advances are made in all areas. All the more reason for educational researchers from a broad array of disciplines, ranging from developmental psychology to psychometrics to natural language processing, to work collaboratively and consider both innovations and limitations in each discipline (for examples, see Sabatini, Albro & O’Reilly, 2012; Sabatini, Albro, & O’Reilly, 2012). Many of the fundamental tools are in place, thanks in large part to the fast pace of technological advances in the cognitive and learning sciences. If properly coordinated, the result is a powerful assessment system that serves equally well as an instructional design mechanism for reading and writing classrooms.

References

Attali, Y. & Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment,* 4. Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1650/1492.

Attali, Y. & Powers, D. (2009). *A developmental writing scale.* (ETS Research Report No. RR-08-19). Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-08-19.pdf

Bennett, R. E. (June, 2011a). *Theory of action and educational assessment*. Paper presented at the National Conference on Student Assessment. Orlando, FL.

Bennett, R.E. (2011b). *CBAL: Results From Piloting Innovative K–12 Assessments***.** ETS Research Report No. RR-11-23. Princeton, NJ: Educational Testing Service.

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K-12 Assessment: Integrating Accountability Testing, Formative Assessment and Professional Support. In C. Wyatt-Smith & J. J. Cumming (Eds.), *Educational Assessment in the 21st Century*: Springer.

Benton, S. L., Corkill, A. J., Sharp, J. M., Downey, R. G., & Khramtsova, I. (1995). Knowledge, interest, and narrative writing. *Journal of Educational Psychology, 87,* 66–79.

Burstein, J. & Chodorow, M. (2010). *Progress and new directions in technology for automated essay evaluation.* In Kaplan, R. (Ed.), The Oxford handbook of applied linguistics, 2nd edition (pp. 487–497). Oxford, UK: Oxford University Press

Burstein, J., Chodorow, M. & Leacock, C. (2003). *Criterion: Online essay evaluation: An application for automated evaluation of student essays.* In Proceedings of the fifteenth Annual Conference on Innovative Applications of Artificial Intelligence (pp. 3-10). Acapulco, Mexico: Association for the Advancement of Artificial Intelligence.

Chodorow, M. & Burstein, J. (2004). *Beyond essay length: Evaluating e-rater's performance on TOEFL® Essays* (TOEFL® Research Report No. RR-73, ETS Research Report No. RR-04-04). Princeton, NJ: Educational Testing Service.

Deane, P. (2012). *Using Writing Process and Product Features to Assess Writing Quality and Explore How Those Features Relate to Other Literacy Tasks.* Manuscript under review.

Deane, P., Sabatini, J. & O'Reilly, T. (2011). *English Language Arts literacy framework.* Educational Testing Service, Princeton, NJ. Retrieved 1 31 13 from: http://www.ets.org/s/research/pdf/ela_literacy_framework.pdf

Elliott, N. (2005). On a Scale: A Social History of Writing Assessment in America. New York: Peter Lang.

French, J.W. (1962). *Schools of thought in judging excellence of English themes. Proceedings of Invitational Conference on Testing Procedures*, 1961. ETS Reprint. Princeton, N.J.: ETS.

Gorin, J. S. (2006). Item design with cognition in mind. *Educational Measurement: Issues and Practice*, *25*(4), 21-35.

Gorin, J. S. (2007). *Test construction and diagnostic testing*. In J. P. Leighton & M. J. Gierl, Eds. Cognitive Diagnostic Assessment in Education: Theory and Practice. Cambridge University Press.

Gorin, J. S. (2012). *Assessment as evidential reasoning.* White paper commissioned by The Gordon Commission on the Future of Educational Assessment. Retrieved on 1/25/2013 from http://gordoncommission.org/rsc/pdfs/gorin_assessment_evidential_reasoning.pdf.

Gorin, J. S. & Embretson, S. E. (2012). *Using cognitive psychology to generate items and predict item characteristics.* In M. J. Gierl, & T. M. Haladyna (Eds.) Automatic Item Generation Theory and Practice, (pp. 136-156). NY, NY: Taylor & Francis.

Gorin, J. S., & Svetina, D. (2011). *Test design with higher order cognition in mind.* In G. Schraw & D. H. Robinson (Eds.), Assessment of higher order thinking skills. Charlotte, NC: Information Age Publishing.

Gorin, J. S. & Svetina, D. (2012). *Cognitive psychometric models as a tool for reading assessment engineering.* In J. Sabatini, E. Albro, and T. O'Reilly (Eds.) Reaching an understanding: Innovations in how we view reading assessment (pp. 169 - 184). Lanham, MD: R&L Education.

Graff, G. (2003). *Clueless in academe: how schooling obscures the life of the mind.* New Haven, CT: Yale University Press.

Hillocks, G., Jr. (2010). Teaching argument for critical thinking and writing: An introduction. *English Journal, 99*(6), 24-32.

Hillocks, G., Jr. (2011). *Teaching argument writing: Supporting claims with relevant evidence and clear reasoning.* Portsmouth, NH: Heinemann.

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication,* 41,201-13.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258-272.

Kuhn, D. (2005). *Education for thinking.* Cambridge, MA: Harvard University Press.

Kuhn, M.R., Schwanenflugel, P.J., Meisinger, E.B., Levy, B.A., and Rasinski, T.V. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly*, Vol. 45, 2, 230-251.

Leighton, J. P., & Gierl, M. J. (2011). *The learning sciences in educational assessment: The role of cognitive models.* Cambridge, UK: Cambridge University Press.

Logan, G.D. (1997). Automaticity and reading: Perspectives from the instance theory of automatization. *Reading & Writing Quarterly*, 13, 2, 123-146. doi:10.1080/1057356970130203.

Jensen, F. V. (1996). *An Introduction to Bayesian Networks*. London: UCL Press.

Jensen, F. V. (2001). *Bayesian Networks and Decision Graphs.* New York : IEEE Computer Society. London: Chapman & Hall.

McCrudden, M. T., Magliano, J., & Schraw, G. (Eds). (2011). *Text relevance and learning from text.* Greenwich, CT: Information Age Publishing.

McCutchen, D. (2006). *Cognitive factors in the development of children's writing*. In MacArthur, D.A., Graham, S., and Fitzgerald, J. (Eds.), Handbook of Writing Research, pp. 115-130. New York and London: Guilford.

McNamara, D.S., Crossley, S.A., and McCarthy, P.M. (2010). Linguistics features of writing quality. *Written Communication, 27(1),* 57-86.

Mislevy, R.J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59,* 439-483.

Mislevy, R. J. & Levy, R. (2007). Bayesian psychometric modeling from an evidence-centered design perspective. In C. R. Rao and S. Sinharay (Eds.), *Handbook of statistics, Volume 26* (pp. 839-865). North-Holland: Elsevier.

Newell, G. E., Beach, R., Smith, J., & VanDerHeide, J. (2011). Teaching and learning argumentative reading and writing: A review of research. *Reading Research Quarterly, 46*(3), 273-304.

O'Reilly, T., & McNamara, D. S. (2007). Reversing the reverse cohesion effect: good texts can be better for strategic, high-knowledge readers. *Discourse Processes 43*(2), 121-152.

O'Reilly, T., Sabatini, J., Bruce, K., & Halderman, L. (2012, July). *Integrating background knowledge measures into high school reading assessment: opportunities and challenges* Paper presented at the Society for Text and Discourse, Montreal, QC.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Pearl, J. (1988). Probabilistic reasoning in intelligent systems: Networks of plausible inference. San Mateo, CA: Kaufmann.

Powers, D. E., Burstein, J., Chodorow, M., Fowles, M. E. & Kukich, K. (2001). *Stumping e-rater: Challenging the validity of automated essay scoring* (GRE® Board Professional Report No. 98-08bP, ETS Research Report No. RR-01-03). Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-01-03-Powers.pdf

Quinlan, T., Higgins, D. & Wolf, S. (2009). *Evaluating the construct coverage of the e-rater scoring engine* (ETS Research Report No. RR-09-01). Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-09-01.pdf

Rupp, A. A. (2012). *Psychological vs. psychometric dimensionality in reading assessment.* In J. Sabatini, E. R. Albro, & T. O'Reilly (Eds.), Measuring up: Advances in how we assess reading ability (pp. 135-152). New York: Rowan & Littlefield Education.

Rupp, A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing, 23,* 441–474.

Sabatini, J., Albro, E., & O’Reilly, T. (Eds.). (2012). *Measuring up: Advances in how we assess reading ability*. Lanham, MD: R&L Education.

Sabatini, J., Albro, E., & O’Reilly, T. (Eds.). (2012). *Reaching an understanding: Innovations in how we view reading assessment*. Lanham, MD: R&L Education.

Shanahan, T. (2006). *Relations among oral language, reading, and writing development.* In MacArthur, D.A., Graham, S., and Fitzgerald, J. (Eds.), Handbook of Writing Research, pp. 171-186. New York and London: Guilford.

Shapiro, A. M. (2004). How including prior knowledge as a subject variable may change outcomes of learning research. *American Educational Research Journal, 41*, 159–189.

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation: Current applications and new directions.* Routledge Academic.

Spandel, V. (2004). *Creating writers through 6-trait writing assessment and instruction.* Boston: Pearson.

U.S. Department of Education (2009). *Race to the Top Program Executive Summary*. U.S. Department of Education, Washington, D.C. retrieved on 1 31 13 from: http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

Wilson, M. & Moore, S. (2012). *An explanative modeling approach to measurement of reading comprehension.* In J. Sabatini, E. Albro, and T. O’Reilly (Eds.) Reaching an understanding: Innovations in how we view reading assessment (pp. 147 – 168).

[i] The framework is consistent with a broader initiative that connects assessment explicitly with best practices in instruction and what is known about student learning and development from the cognitive and learning sciences literatures. This initiative is termed ‘Cognitively Based Assessment of, for, and as Learning,’ or CBAL for short (Bennett, 2011; Bennett & Gitomer, 2009).