

Final Paper

The Hannon and Daneman (2001) Reading Comprehension Assessment: A Critique from an ECD Perspective

Introduction

Reading comprehension assessments, as Brenda Hannon and Meredyth Daneman (2001) argue, have historically lacked a basis in psychological models of reading. Reading tasks seem to have been developed arbitrarily, without a design for tapping different aspects or components of comprehension. Therefore, their argument for the relationship between the evidence—performance on the assessment—and the claim that an examinee can comprehend (cf. Schum, 1994) is weak. Hannon and Daneman therefore set out to develop a new theory-based assessment for reading comprehension in college students, and made an intriguing finding; their assessment accounted for 60-65% of the variance in standardized Nelson-Denny reading test scores.¹

¹ Note that that this explains variance between examinees, but variance within examinees might be interesting, especially in an individual differences approach.

But how well do the different aspects of their assessment stand up to their stated goal of developing a theory-based assessment, and how internally consistent is the assessment? That is, how good is their argument that performance by the target population on their assessment is strong and clear evidence for claims about true reading comprehension? I will argue that while relationships among many aspects of their assessment are strong (e.g., their task model is a strong fit with their student model), their psychological model is not a good fit with either mainstream psychological models of reading comprehension, or with the literature that they say they are basing their assessment on; and this is a fatal flaw in the evidentiary chain of reasoning. I will argue that while Hannon and Daneman's assessment does provide evidence in support of a claim about performance on their task and memory, it does not provide evidence in support of claims about examinees' reading comprehension, at least not as comprehension is usually understood by reading researchers.

I will begin by describing the assessment and its stated purpose, especially as that relates to psychological theories of reading comprehension (§1). I then discuss the student model in terms of both psychological theory (§2) and purpose (§3). In §4-§6 I discuss the statistical model with regard to the student model, theory, and purpose. I then turn to the task model in §10-13, and finally the evidence rules in §14-18. The entire critique is summarized in Appendix A.

Description of the Assessment

Hannon and Daneman have developed a reading comprehension assessment that taps accessing prior knowledge, making inferences from the text, integrating accessed knowledge with text information, and recalling what was read. The assessment consists of 6 three-sentence passages, with each passage followed by 42-54 questions. Passages include both real and nonsense nouns, connected by 2-4 types of relationships (e.g., height, weight, color, size). A sample passage is: "A NORT resembles a JET but is faster and weighs more. A BERL resembles

a CAR but is slower and weighs more. A SAMP resembles a BERL but is slower and weighs more” (p. 109). In order to answer the questions, examinees must access from prior knowledge that fact that a jet is both faster and heavier than a car. They must also make inferences from items in the text (e.g., Speed: BERL < CAR, SAMP < BERL therefore SAMP < CAR). And they must integrate prior knowledge with text information (another form of inference; e.g., For weight, from prior knowledge—JET > CAR; from the text—NORT > JET; inference—therefore NORT > CAR).

Hannon and Daneman found their assessment explained 60-65% of the variance in scores on a standardized reading comprehension test (the Nelson-Denny), was quick to administer (approximately 30 min.), and provided separate scores for knowledge access, integration, inference, and text memory.

Purpose

§1 The stated purpose of the assessment is not for practical use in academic settings (either to measure schools' or students' progress in reading skills or to diagnose reading problems), but for psychological theory building. As such, the authors have a particularly strong responsibility to ground their assessment well in existing theories.

Psychological Basis. Hannon and Daneman base the design of their reading comprehension assessment on what they term a multi-component (cognitive psychology) reading comprehension theory, combining semantic, syntactic, and referential relationships. They distinguish this theory from both word-level and lexical access theories of comprehension and from higher-level process theories that draw on only a single component. I will argue that Hannon and Daneman have been only partially successful in identifying “what knowledge, skills, or attributes” should be assessed (Messick, 1994, p. 16), partly because they have not consistently applied the research they draw on, but also because they have a narrow reading of the research with regard to the inferences they wish to draw.

Hannon and Daneman present evidence that once students have mastered early reading, higher-level processes account for most of the variance in comprehension (e.g., Jackson & McClelland, 1979). Furthermore, they argue that it is important to include a number of components, given the evidence that several higher-level processes make independent contributions to comprehension; prior knowledge (Haenggi & Perfetti, 1994); vocabulary (Dixon, LeFevre, & Twilley, 1988); working memory (Dixon, LeFevre, & Twilley, 1988; Haenggi & Perfetti, 1994); integrating information from prior knowledge (Dixon, LeFevre, & Twilley, 1988); and monitoring and repairing inconsistencies (Paris & Lindauer, 1976). They conclude by stating that current multi-component theories require assessments that tap multiple components, but that no single theory-based multi-component test exists. (cf. Pellegrino,

Chudowsky & Glaser, 2001 for the trend in cognitive psychology for multi-component models, and the need for them in assessment).

However, having laid out other researchers' multi-component models, Hannon and Daneman fail to articulate the specific components of their own multi-component model. The observed variables are memory for just-read text (which they call text memory) and various inferencing processes (both within-text and between prior knowledge and text, which they call text inferencing). Inferencing draws heavily on working memory, so the assessment implicitly taps working memory. They explicitly exclude background knowledge from their test by creating tasks that only draw on facts known by their target undergraduate population (e.g., an ostrich has a long neck). Hannon and Daneman's multi-component model therefore omits several important higher-level processes that they themselves identified as important for reading comprehension; vocabulary, prior knowledge, and a wide range of strategies for monitoring and repairing inconsistencies.

With regard to vocabulary, they argue that their assessment accounts for the role of vocabulary in reading comprehension, since vocabulary scores did not add significantly to Nelson-Denny scores once they had partialled out knowledge inferencing, knowledge integration, knowledge access, and speed. This seems to be a weak argument for two reasons: 1) vocabulary is a form of knowledge, and 2) the Hannon and Daneman test uses a restricted vocabulary, partly because it is so short. Although it is laudable to try to control construct-irrelevant variance, Hannon and Daneman here attempt to control construct-*relevant* variance—a person who can only comprehend text written with a small vocabulary is like a musician who can play beautifully . . . but only in one key! With regard to working memory, Hannon and Daneman's assessment does correlate highly with measures of working memory (not surprisingly, since the authors' comprehension research to date has focused on working memory; e.g., Daneman &

Carpenter, 1980).

The Claim. The inference that Hannon and Daneman want to make about examinees (the claim) is that they can comprehend what they read. Hannon and Daneman seem to have a conventional view of what comprehension, as an end product, looks like. That is, a comprehender can read and then correctly answer questions of various types from short passages of an academic nature. For example, Hannon and Daneman compare college students' performance on their assessment to performance on the Nelson-Denny comprehension test. The choice of the Nelson-Denny test is an interesting one, since it uses only narrative passages, while the Hannon and Daneman passages are more like (a very odd sort of) expository text.

Hannon and Daneman suggest the test be used as a research tool and for theory-building. However, their claim is not situated in any context or purpose for reading (e.g., everyday uses of literacy in the YALD, Sheehan & Mislevy, 1990; academic uses of literacy in the ACTFL or college uses of language in the TOEFL, Mislevy, 1994; Mislevy, Steinberg & Almond, in press-a, in press-b; see Appendix B), learning goals, type of text, or genre. Note that current theories of inference in reading narrative text *do* differentiate between types of text and reader purposes (Graesser, Singer & Trabasso, 1994).

Commentary on Hannon and Daneman's Psychological Foundation

The reading comprehension literature is vast, replete with methodological problems, weak on theory, and strong on intuition (National Reading Panel, 2000; Shanahan, 2000). That said, Hannon and Daneman have failed to build on the most widely cited theories of reading comprehension (e.g., Just & Carpenter, 1992; Kintsch, 1988, 1998; Kintsch & van Dijk, 1978; LaBerge & Samuels, 1974; Perfetti, 1985).

Any theory in reading must account for a vast body of findings. Graesser and Britton (1996), point out that a unified cognitive theory of reading comprehension (which does not yet

exist) needs to account for the experimental results in recall, summarization, question answering, ratings of relatedness, reading times, naming latencies, and think-aloud protocols and must account for different types of texts, genres, and reader goals. Such a theory, Graesser and Britton argue, is likely to:

- a) include a surface code, propositional textbase, and situation model which
- b) together, in a complex dynamical system form a coherent representation,
- c) using inference processes within text and from background knowledge, and
- d) within the constraints of working memory.

Clearly, Hannon and Daneman's model of reading comprehension omits major parts of this theory and therefore their assessment fails to tap into components that are needed to support their claim that their assessment measures reading comprehension.

Student Model

Identified.

The Hannon and Daneman (2001) assessment uses a 4-part student model, with 2 main areas—*inference* and *memory*. Inferences include *text inferencing* (TI—making inferences within what was read) and *knowledge integration* (KI—making inferences between prior knowledge and what was read). Memory includes *text memory* (TM—remembering what was read), and *knowledge access* (KA—access to prior knowledge). The authors report all 4 parts of the student model when they report results of the assessment. They explicitly argue, consistent with their purpose of theory-building, for an individual differences approach, that is, there is no single examinee score, but 4 separate scores (TI, KI, TM, and KA). They also further break down knowledge integration into low and high levels, and knowledge access components into low, medium, and high levels.

Despite the argument about individual differences, Hannon and Daneman report

relatively high intercorrelations among the 4 components ($r = .17$ to $.83$ across 40 correlations, 29 are $> .50$). Also, when the authors use the assessment for prediction in a regression model, they seem to collapse all 4 parts into one, suggesting that perhaps they really envision a 1-part student model with 4 subparts.

Relation to the Psychological Model

§2 The inference and memory variables in the Hannon and Daneman assessment represent some, but not all, of the salient variables that the authors identified among multi-component models of reading comprehension. Inference is prominent in the literature that Hannon and Daneman reviewed, whereas memory is not particularly prominent. As pointed out above, a major flaw in the assessment is the omission of several important components of comprehension, such as background knowledge, vocabulary, and the use of cognitive and metacognitive strategies.

Relation to the Purpose of the Assessment

§3 Given that Hannon and Daneman's purpose for the assessment is theory building, their student model is suited to building their particular theory of comprehension (i.e., comprehension = memory + inference), but not to commonly accepted theories of comprehension.

Evidence Model—Statistical Model

Statistical Model

Although Hannon and Daneman separate the 4 components when they report scores, their review of the literature on multicomponent models of reading comprehension suggests that they take something like a classical test theory approach—a good reader tends to answer many TI, KI, TM, and KA questions correctly, and a poor reader tends to answer all four incorrectly. Those

answers reflect some “true” score on each of the 4 types of questions for the reader.

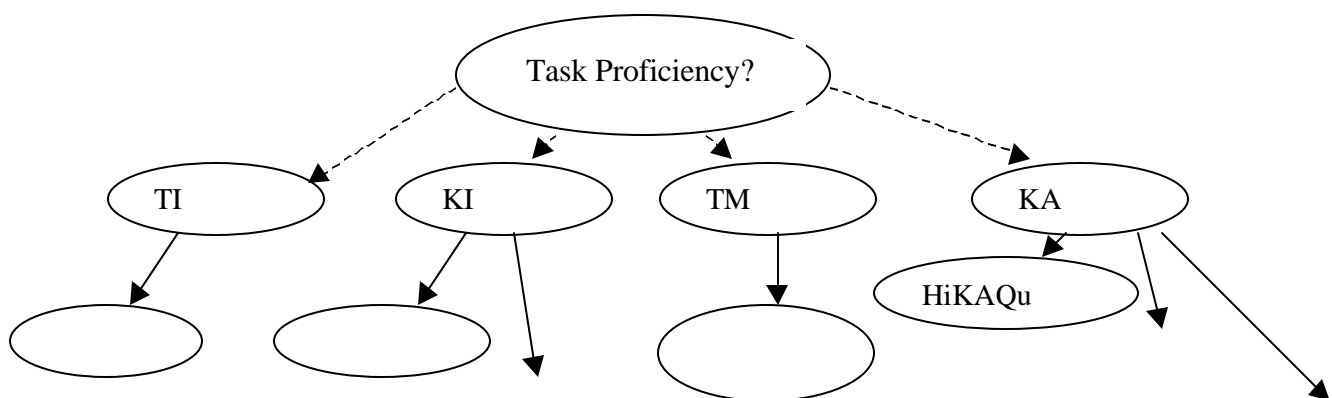
At the same time, Hannon and Daneman’s individual differences approach seems to leave open the possibility that an examinee could tend to answer one type of question right and the other types wrong, which would be an interesting pattern for theory-building. In addition, as mentioned above, when they regress all 4 components on Nelson-Denny comprehension, they seem to suggest a 1-variable “task proficiency” student model, which has implications for the statistical model. These questions remain to be tested empirically, in which case the description of the student model might need to be modified.

Observable Variables

§4 Observable variables for this assessment consist of True/False answers to questions about the text (“A NORT is faster than a JET.” True or False?). These are at varying levels of difficulty (e.g., knowledge integration, low and high). Observable variables are therefore the answers to specific types of questions presented to examinees. The work product, therefore, is very simple and has little of the richness of the cognitive psychology of reading that their theory drew on (e.g., studies using retelling, application questions, verbal protocols, etc.). In summary, their observable variables do not tap many aspects of reading comprehension theory; but what they do tap, they tap well.

Diagram of Student Model and Observable Variables

The diagram below shows the relationship between the student model and the observable variables.



TIQu

HiKIQu

TMQu

LoKIQu

MedKAQu

LoKAQu

Note: TIQu = TI questions.

Note that I have shown KI as a student model variable, with LowKI and HighKI as observable variables. But it is not clear whether Hannon and Daneman mean for LowKI and HighKI to be conditionally independent, which the diagram implies (correlations, as noted above, are high). The same caveat applies to KA. Again, these questions remain to be tested empirically, and the description of the student model might need to be modified.

.§5-6 With regard to the purpose of the assessment—psychological theory building—the observable variables fit Hannon and Daneman’s restricted theory of comprehension (i.e., comprehension = memory + inference), but not mainstream comprehension theories (ones that also include vocabulary, prior knowledge, and cognitive and metacognitive strategies). Overall, Hannon and Daneman’s statistical model, like their student model, is not a very good fit with the research base they say they are basing their assessment on.

Task Model

Description of the Task Model—Tasks

The tasks in Hannon and Daneman’s assessment consist of *passages* that examinees read and true/false *questions* that they later answer. The tasks are highly structured; so much so that they *could* easily be generated by an algorithm (cf. Embretson, 1998), although the authors do not seem to have done so.

Fixed Features

Tasks. The assessment consists of 1 practice passage with questions and 6 test passages with questions.

Passages. Each passage consists of 3 sentences, each containing 2 terms and 1 semantic feature. In each passage a total of 2 real terms and 3 nonsense terms are used. All real terms and semantic features are selected from those known to college students (the population that the assessment was designed for and piloted on). Although not specified, each sentence begins with a nonsense word. Although not explicitly stated, all semantic features are real words, not nonsense words.

Sample passage:

[TERM] [TERM] [semantic features]
“A TOLP resembles a MARB but is more colorful, larger and lives in a colony.
A MARB resembles a BUTTERFLY but is more colorful and larger.
A JERP resembles an ANT but is less colorful and larger.” (p. 126)

In the example above, TOLP, MARB, and JERP are the nonsense terms, BUTTERFLY and ANT are the real terms, and “colorful, larger” and “lives in a colony” are the semantic features.

Questions. All questions are true/false statements. One half of the statements are true, and one half are false, created by reversing the order of the terms in a correct statement.

Sample questions:

TM: “A MARB is more colorful than a BUTTERFLY.” [T]
TI: A BUTTERFLY is more colorful than a TOLP. [F]
KI (low): “A MARB is larger than an ANT.” [T]
KA (high): COCKROACHES have queens, whereas ANTS don’t. [F] (p. 126-127)

Variable Features

Tasks. The number of features per paragraph is 2 in paragraphs one and two, 3 in paragraphs three and four, and 4 in paragraphs five and six.

Passages. Each passage uses 2-4 semantic features; there does not seem to be any order

in which they are repeated across the passage. When there are more than 2 features, which sentence has 3 features also varies from passage to passage, in no apparent order. Among the 2 real terms and 3 nonsense terms, 1 term is reused across the 3 sentences, but it can be either a real or nonsense term. Presumably, the authors believe that these variations do not introduce construct-irrelevant variance.

Questions. The design and number of questions for each type of question is complicated, and is represented in the table below. Each 3-sentence passage has 42-54 questions.

	Level	Is it used in the text?			Number of questions	
		Term	Term	Semantic feature	If this many semantic features are used	...there are this many questions (½ true, ½ false)
Text memory	—	Y	Y	Y	2	12
					3	14
					4	16
Text inferencing	—	Y	Y	Y	2	4
					3	6
					4	8
Knowledge Integration	Low	Y	Y	Y	Any number	4
	Middle	N	Y	Y	Any number	6
	High	N	Y (nonsense)	N	Any number	8
Knowledge Access	Low	Y	Y	Y	2	4
					3	6
					4	8
	High	Y	N	N	Any number	4

For example, a passage with 2 semantic features has 42 questions: 12 TM questions + 4 TI questions + 4 KI (low) questions + 6 KI (middle) questions + 8 KI (high) questions + 4 KA (low) questions + 4 KA (high) questions.

Although not specified, questions follow one of 4 formats:

- 1) A [TERM] is/does [semantic rel].
- 2) A [TERM1] is/does [semantic rel] than a [TERM2].
- 3) Like [TERM1], [TERM2] is/can [semantic rel]—used for KI(high) questions.
- 4) A [TERM1] has/is [semantic rel], whereas a [TERM2] doesn't/isn't—used for KA(high) questions.

Conditions of Administration

Each passage is presented on a computer screen 1 sentence at a time in a fixed order, and is replaced when the examinee pushes the “+” key. Each question is presented 1 at a time in random order, for a maximum of 12 sec per question. Examinees may not scroll backwards or forwards, either within the passages, within the questions, or between the questions and passage. Hannon and Daneman assume, but do not provide evidence, that this increases the assessment’s ability to measure their version of comprehension.

Stimulus Material and Work Product Specifications

Stimulus materials are presented on a computer screen; there are no specifications regarding the font size or color, background color, screen size, or programming language. The real and nonsense terms are always presented in CAPITAL letters, but the semantic relations are in lowercase letters. The “work product” is a key press on the computer keyboard signifying Yes for True or No for False. Failure to answer is counted as an error. There are no specifications regarding criteria for key presses (e.g., how to interpret an ambiguous key press).

Rationale—Relationship of the Task Model to Other Aspects of the Assessment

§7 With respect to the student model, each question is designed to tap one (and only one) of the student model variables, with questions differentiated among low, middle, and high levels of student model variables. (This suggests a 7-variable student model as opposed to a 4-variable student model, but again this needs to be tested empirically, and perhaps the student model

modified.) The TM, TI, KI, and KA formats are clearly linked to the student model. Hannon and Daneman do use the literature on text memory and text inferencing to specify what makes a questions easy or hard to answer for a particular passage (e.g., the number of relationships, amount of inference required, amount of access to prior knowledge required). Although they did not design the original assessment, they did adapt it specifically in order to tap text memory and text inferencing.

However, it is open to doubt whether each question taps only one student model variable—text memory (as our ETS counterparts pointed out in the April 29th class) is surely involved in all of the questions. Further, while the *design* of the questions seems linked to the student model, the *number* of questions does not. That is, what is it about TM that demands 12 questions, not more or less, and what is it about and KI(middle) that demands 6 questions, not more or less? Likewise, what is it about the student model that dictates, for example, that nonsense words come first or that only KI(high) questions can take the format shown in example 3 above? Furthermore, counting non-answers (unable to solve the problem) the same as wrong answers (able to solve, but incorrectly) fails to differentiate between situations that are probably different.

Finally, the true/false answer format is a crude and noisy format. It misses the complexity of reading shown by think-aloud protocol studies of reading comprehension (e.g., Haas & Flower, 1988). Finer gradations of answers (e.g., most right/somewhat right/wrong) could strengthen the relationship between the task model and the psychological model.

§8 With respect to the statistical model, each answer has only 1 observable variable (unlike many performance assessments (e.g., Clauser et al., 1997). The answer provides the True/False evidence required by the evidence interpretation model and what appears to be a classical test theory model.

§9 With respect to the psychological model, it is unclear—either in the literature or in this assessment—what the interaction among the observable variables (whether they are conditionally independent) might be. If the observable variables are not conditionally independent, that has implications for the relationships among the task design specifications that produce each observable variable. A more serious problem is that the tasks clearly fail to tap the vital comprehension components of vocabulary and world knowledge that Hannon and Daneman cite in their own literature review.

§10 With respect to the purpose of the assessment—psychological theory building—again the task design fits Hannon and Daneman’s restricted theory of comprehension, but not mainstream comprehension theories.

Overall, the task model has the strongest fit with the student model and the purpose, a somewhat weaker fit with the statistical model, and the weakest fit with the psychological model.

Evidence Model—Evidence Rules

The evidence rules for this assessment are very simple, since the work product is simply a true or false answer. Unlike a performance assessment, with multiple raters, complex work products, and rubrics that address many aspects of each work product (e.g., Clauser et al., 1997), answers are simple to score.

§11-15 With regard to the task model, even though questions tap different aspects of comprehension, each answer is of the same type. With regard to the statistical model, the answer to each question provides one piece of evidence specific to the score in the CTT model. Given the task the authors have designed (one student model variable per question, with a T/F answer), there is a good fit overall of the evidence rules with the other components. Producing a quick, easy-to-score assessment was one stated goal for these authors, and having a simple answer

format and simple evidence rules helps them accomplish this.

Conclusion

Other Critiques of Reading Comprehension Assessment

Many critics, both reading researchers and practitioners, have found fault with existing standardized reading assessments. Hannon and Daneman's assessment should perhaps address some of these failings of existing assessments. These criticisms are rooted in practical, sociocultural, and traditional measurement theory (e.g., validity) perspectives. From a practical perspective, critics find fault with large-scale (e.g., national, state-wide, or district-wide) assessments for failing to provide teachers with information that is useful to them in assisting individual students (e.g., diagnostic information) (see Leipzig & Afflerbach, 2000). However, it seems that these critics fail to recognize that tests are designed to fulfill one purpose, and cannot be used for another purpose without being adapted (Almond, Steinberg, & Mislavy, 1999).

From a sociocultural perspective, critics have charged that standardized reading tests are incapable of assessing the rich, situated, social and interactive nature of the reading task (see Murphy, Shannon, Johnston, & Hansen, 1998)—essentially a situative perspective (see Pellegrino et al., 2001). Perhaps these critics, too, misunderstand what the test designers set out to measure—a particular kind of academic literacy, which places its own unique demands on vocabulary, syntax, knowledge, and pragmatics (Sheehan & Mislavy, 1990). On the other hand, reading assessment authors should be clear about what they mean when they make the claim that a student who has passed their test “can read.” That is, the test shows the ability to use reading to perform a certain (rich, situated, social and interactive) type of reading—academic reading. The same student may or may not be able to read other types of texts, about other types of subject matter, for other purposes, in other settings (See Appendix A).

From a traditional measurement theory perspective, reading comprehension assessments

have been criticized on validity grounds. These validity criticisms are related to the sociocultural criticisms—if the challenges in a test’s reading passages do not match the challenges of “real reading,” then children’s performance on the test is not good evidence of their “real” reading ability (see Hill & Larsen, 2000). Perhaps this is what teachers react to when they give an assessment low marks for face validity. The high concurrent validity of reading comprehension assessments has sometimes been criticized, on the grounds that the other standardized tests that a particular assessment is being compared to are invalid to begin with (Murphy et al., 1998).

A final basis for criticizing existing reading comprehension assessments is their low construct validity or relationship to theory. Hannon and Daneman themselves put this forth as a justification for having developed their assessment.

Summary

The biggest weaknesses of Hannon and Daneman’s assessment from an ECD perspective are the disconnect between a) the multiple components that we know from psychology research to make up comprehension, b) the small number of student model variables, and c) the crudeness of the true/false format (part of the task model) for measuring those student model variables. This assessment could be an interesting first step towards building such a multi-component assessment, but I do not believe that was the authors’ intention—they do believe that text memory and text inferencing *are* reading comprehension.

Other than these foundational flaws, however, the assessment is relatively strong on the relationships among the purpose, student, evidence, and task models. The simple evidence rules required for a true/false format are a good fit with the other components. The three models fit well with the purpose of the assessment. The statistical model and task model are generally a good fit with the other components, with one major caveat: theory has not well specified the interrelationships of various components of comprehension (i.e., student model variables), and

neither does this assessment. For example, it is not clear whether the observable variables LowKI and HighKI are conditionally independent with respect to the student model variable KI (although we know them to be highly correlated).

Implications for Future Reading Assessments

An ideal assessment would have a clearly stated purpose, would have student model variables that were consistent with a current theory or theories of reading comprehension, and, like Hannon and Daneman's assessment, would have a task model and statistical model that were consistent with the student model and the psychological theory. As Hannon and Daneman themselves point out, current assessments are not theory-based, but ad hoc. Although some of them have high predictive validity (e.g., SAT verbal scores and first-year success in college), it is not because they were *designed* to tap important aspects of comprehension, but because they *happen* to.

Future assessments should draw on a number of components of comprehension, including background knowledge, cognitive and metacognitive strategies, inference, vocabulary, and working memory. The program for developing such an assessment might begin, not with all of the components, but with selected subsets of components, in order to tease out the contribution of various components to performance. Such an assessment might then be useful not only for accountability purposes, but might also be modified for research, diagnosis, and/or instruction.

Future assessments, their purpose and psychological basis will need to be explained to the many audiences for reading tests: school personnel, parents, and policymakers. One lesson, I think, from the current controversy over the MSPAP is that the purpose of an assessment is often misunderstood, which leads to unwarranted criticisms. Clearly specifying and communicating the purpose would then increase face validity and alleviate some of the practical obstacles that assessments have faced.

References Cited

- Almond, R.G., Steinberg, L.S., & Mislevy, R.J. (1999). *A sample assessment using the four process framework*. White paper prepared for the IMS Working Group on Question and Test Inter-Operability. Princeton, NJ: Educational Testing Service.
- Clauser, B.E., Ross, L.P., Clyman, S.G., Rose, K.M., Margolis, M.J., Nungester, R.J., Piemme, T.E., Chang, L., El-Bayoumi, G., Malakoff, G.L., & Pincetl, P.S. (1997). Development of a scoring algorithm to replace expert rating for scoring a complex performance-based assessment. *Applied Measurement in Education, 10*, 345-358.
- Daneman, M., & Carpenter, P.A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*, 450-466.
- Dixon, P., LeFevre, J.-A., & Twilley, L. (1988). Word knowledge and working memory as predictors of reading skill. *Journal of Educational Psychology, 80*(4), 465-472.
- Embretson, S.E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*, 380-396.
- Graesser, A., & Britton, B. K. (1996). Five metaphors for text understanding. In A. Graesser & B. K. Britton (Eds.), *Models of understanding text* (pp. 341-351). Mahwah, NJ: Erlbaum.
- Graesser, A., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*(3), 371-395.
- Haas, C., & Flower, L. (1988). Rhetorical reading strategies and the construction of meaning. *College Composition and Communication, 39*(2), 167-183.
- Haenggi, D., & Perfetti, C. A. (1994). Processing components of college-level reading comprehension. *Discourse Processes, 17*, 83-104.

- Hannon, B. & Daneman, M. (2001). A new tool for measuring and understanding individual differences in the component processes of reading comprehension. *Journal of Educational Psychology, 93*(1), 103-128.
- Hill, C., & Larsen, E. (2000). *Children and reading tests*. Stamford, CT: Ablex.
- Jackson, M. D., & McClelland, J. L. (1979). Processing determinants of reading speed. *Journal of Experimental Psychology: General, 108*(2), 151-181.
- Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*(1), 122-149.
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*(2), 163-182.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, England: Cambridge University Press.
- Kintsch, W., & van Dijk, T. A. (1978). Towards a model of text comprehension and production. *Psychological Review, 85*, 363-394.
- LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6*, 293-323.
- Leipzig, D. H., & Afflerbach, P. (2000). Determining the suitability of assessments: Using the CURRV framework. In L. Baker, M. J., Dreher, & J. T. Guthrie (Eds.), *Engaging young readers: Promoting achievement and motivation* (pp. 159-187).
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.
- Mislevy, R.J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*, 439-483.
- Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (in press-a). On the structure of educational

- assessments. *Measurement*.
- Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (in press-b). Task-based language assessment. *Language Assessment*.
- Murphy, S., Shannon, P., Johnston, P., & Hansen, J. (1998). *Fragile evidence: A critique of reading assessment*. Mahwah, NJ: Erlbaum.
- National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: National Institute for Child Health and Human Development.
- Paris, S., & Lindauer, B. K. (1976). The role of inference in children's comprehension and memory for sentences. *Cognitive Psychology*, 8, 217-227.
- Pellegrino, J.W., Chudowsky, N., & Glaser, R. (Eds.) (2001). Advances in the sciences of thinking and learning. Chapter 3 in *Knowing what students know: The science and design of educational assessment* (pp. 58-109). Washington, DC: National Academy Press.
- Perfetti, C. (1985). *Reading ability*. NY: Oxford University Press.
- Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning*. NY: Wiley.
- Shanahan, T. (2000). Research synthesis: Making sense of the accumulation of knowledge in reading. In M. L. Kamil, P.B. Mosenthal, P.D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. 3). Mahwah, NJ: Erlbaum.
- Sheehan, K.M., & Mislevy, R.J. (1990). Integrating cognitive and psychometric models in a measure of document literacy. *Journal of Educational Measurement*, 27, 255-272.

Appendix A

	Purpose	Research/ Theory	Student model	Evidence model— Stat model	Task model	Evidence model— Evidence rules
Purpose	—	§1 Purpose is theory building, but the assessment does not tap many aspects of existing theory	§3 Purpose is theory building, S. model is suited to that	§6 Purpose is theory building, observable variables are not well differentiated for that	§10 Purpose is theory building, task model is suited to that	§15 Purpose is theory building, CTT evidence rules are suited to that
Research/ Theory		—	§2 S. model does not tap many aspects of theory; what it does tap, it taps well	§5 Observable variables do not tap many aspects of theory; what they do tap, they tap well. Interaction of components (e.g., conditional independence) is not well established in theory or in this assessment	§9 True/false answer format misses the complexity of reading shown by think-alouds	§14 True/false requires very simple evidence rules
Student model			—	§4 Observable variables are a good fit with student model, OV's differentiate slightly more than SMV's do. Interaction of OV's (e.g., conditional independence) is not well established in theory or in this assessment	§8 Task model an excellent fit with 4 student model variables—each question is built to tap 1 student model variable. True/false answer format is an OK fit with the 4 student model variables—could have differentiated more	§13 True/false fits with student model
Evidence model— Stat model				—	§7 Task model an excellent fit with observable variables—each question is built to tap 1 student observable variable. Interaction of OV's (e.g., conditional independence) is not well established in theory or in this assessment	§12 True/false feeds into CTT model
Task model					—	§11 True/false answer format an OK fit with OV-specific questions, could have differentiated more
Evidence Rules						—

