13.

Matters of Goodness

Knowing and Doing Well in the Assessment of Critical Thinking

Sharon Murphy

To assess critical thinking is to comment on its goodness. Yet, as the chapters in this book reveal, the essential goodness of critical thinking is a complex and highly contentious matter. Engaging in the educational assessment of critical thinking compounds questions of goodness as it raises questions about not only the goodness of critical thinking, but also about the goodness of the assessment methodology one employs.

Given these challenges, how does one conduct oneself well in the assessment of critical thinking? In answering this question, I draw upon the philosophical exploration of "epistemic responsibility" articulated in Code (1987). For reasons I discuss, I believe it can help in the development of assessment strategies, tools, and practices that can inform the teaching of critical thinking.

Epistemic Responsibility, Assessment, and Critical Thinking

Code (1987) begins with the assumption that "most so-called knowledge is really well-warranted belief" (47). As Fleck (1979) and other philosophers of science note, yesterday's flat earth (a "fact" well warranted by the arguments during the time of that belief) is today's object of benevolent amusement and even ridicule. Situating knowledge claims in this way suggests that facts do not stand on their own and need to be understood within the context in which they occur.

The first step in justifying knowledge claims is to provide reasonable argumentation to sustain the believability, or goodness, of the claims made within a particular field. But it is a mistake to think that this is all there is to justification. According to Code, **epistemic responsibility** is tied to moral responsibility. As she puts it, “in some sense, ethical responsibility is founded upon epistemic responsibility, even if it is not identifiable with it” (5), and “one who has not been scrupulous in knowing cannot be scrupulous in doing” (95). It is reasonableness of conduct, not absolute rightness or wrongness, that is the central concern in ethical conduct, and it can be judged only by considering and understanding context. To discharge one’s epistemic and moral responsibility, one must therefore not only be concerned with the usual matters of evidence and justification, but also be sensitive to context.

This means that when we argue about the goodness of our claims, we must simultaneously consider their goodness or lack of goodness in a particular situation in a particular community. In short, the epistemically responsible approach “denies the autonomy of the known, maintaining that the nature of the knower and of his or her environment and epistemic community are epistemologically relevant, for they act as enabling and/ or constraining factors in the growth of knowledge” (Code 1987, 28).[1]

Against this background, one might ask what it would mean to engage in an epistemically responsible assessment of critical thinking. Following Code, this requires that three things be considered: (a) the knowledge claims made about assessment, critical thinking, or both; (b) the manner in which such claims are warranted; and (c) how such claims situate themselves in a particularity of context. One might compare these requirements for epistemic responsibility to the challenges that Messick (1989) offers. Writing in the field of educational and psychological assessment, he argues that consequential validity — the value implications and social consequences of assessment — need to be considered in validity arguments about the goodness of the assessment (see Ennis, this volume).

Conceptions of Critical Thinking?

Claims about **critical thinking** are not in short supply. One popular internet search engine returned over fifteen million hits for "critical thinking." Some comments, such as the following from the actor, Alec Baldwin, bemoan a shortage of critical thinking:

> There's less critical thinking going on in this country on a Main Street level — forget about the media — than ever before. We've never needed people to think more critically than now, and they've taken a big nap. (Brainy Quote web-site)

Others offer seemingly seductive promises:

> A great way to get kids to think is with materials from *Critical Thinking* [Company]. Be amazed, as I was, at the number of quality thinking products. Great, great stuff — and the work is done for you! (Kidsdomain.com, as cited on the Critical Thinking Company website)

Although these short quotes indicate the ordinary person's understanding of the term "critical thinking," they, like the chapters in this volume, also illustrate the differences that characterize attempts to conceptualize critical thinking. In keeping with the requirements of epistemic responsibility, these differences suggest that such claims must be considered in a context that gives them meaning. Context is doubly important in the case of critical thinking because critical thinking is, by definition, relational — it is done in relation to an action, object, person, event, idea, or situation.

In her account of epistemic responsibility, Code argues that the process of determining the warrantability of any claim should include seeking advice from persons knowledgeable about the area in question. When a term like "critical thinking" is used, users have to consider whether the debates that characterize the term reflect different epistemologies or the effects of different contexts. An attempt to tease out the warrantability of each claim made about critical thinking would require a detailed analysis of each situational use — a task that lies beyond the scope of this discussion. However, it is possible to sketch inclusive and exclusive parameters that establish some criteria which are and are not constitutive of critical thinking. In this way a foundation can be created for the evaluation of claims that are made about critical thinking.

One exclusionary criterion for critical thinking that is implicit (and sometimes explicit) in the chapters in this volume holds that critical thinking must not be mistaken for critique. Johnson cautions that the "critical" in critical thinking is not tantamount to "criticism." A particular critique may be an example of critical thinking at work, but not all examples of critical thinking are examples of critique, and not all examples of critique are examples of critical thinking. Critique is not, therefore, a satisfactory criterion for critical thinking.

Similarly, the chapters on creativity (Hare and Sobocan) suggest that although some examples of creative thought may be examples of critical thinking, not all examples of critical thinking are examples of creative thinking. Indeed, some instances of critical thinking may be achieved quite mechanically — as when one follows a set of prescriptive processes to think critically about something. So, again, creative thought is not a satisfactory criterion for determining critical thinking. Likewise, dispositions and commitments (see Giancarlo-Gittens and Case) are stances towards engaging in critical thinking but are not necessary or sufficient determinants of critical thinking.

Taken as a whole, the essays in this book suggest that the exclusionary boundaries for critical thinking are more easily drawn than the inclusionary ones. Nonetheless, two central themes permeate the inclusionary criteria that various authors suggest — argumentation[2] and judgment. These two themes work in tandem: critical thinkers must be familiar with the conceptual bases of a set of ideas, as well as the evidentiary basis behind the ideas so they may assess their merits and, as the situation demands, put forth credible ideas of their own. The demands on the critical thinker are, in essence, to perceive the essential points in a set of ideas, to ideate (categorize, conceptualize, hypothesize and think openly, analyze, generalize, think conditionally) in relation to that set of ideas, and to represent or present (as the situation demands) a response within the same knowledge domain (Goodman, Smith, Meredith, and Goodman 1987, 15). Each of these actions demands incrementally more of the thinker engaged in critical thinking.

Though the experts in this volume generally hold that argumentation and judgment are central to critical thinking, they differ in their individual articulations of what counts as argumentation and what is involved in judgment. In describing the nature of argumentation, van Eemeren and Garssen use discourse analysis as a heuristic device. Johnson is particularly concerned with the dialectical tier — how a discourse may be structured in terms of the anticipatory moves one makes in relation to opposing points of view. Hare's interest is less in contestation and more in openness and analytical thinking. Nosich's emphasis on the fundamental and powerful concepts within disciplines highlights the particularity of argumentation within these disciplines (and, as such, amplifies its contextual elements).

Another central theme unifying the essays in this volume is that education in critical thinking skills is either explicitly (Kaser, Blair, Hatcher) or implicitly both necessary and good. Although it does seem likely that better thinking should enhance one's quality of life and perhaps that of others, the necessity of instruction in critical thinking and the form that instruction should take are more contentious. Some authors endorse an extreme version of the need for instruction in critical thinking skills; Blair, for example, argues that critical thinking is not innate and must be acquired.[3] Others situate critical thinking in practices within specific social contexts such as a democracy, a discipline (Nosich), creative activity (Hare), or the application of deductive logic in writing (Hatcher). Although the thematic focus of the discussion naturally encourages an emphasis on education, non-school occasions that highlight the goodness of critical thinking warrant more attention. Few are likely to deny that there are non-school contexts in which critical thinking skills are alive and well. Even in pre-school years, children pragmatically negotiate their way toward increasing the number of toys they have, postponing their bedtimes, or extending the amount of play time they have by analyzing situations, anticipating the arguments offered by parents, and creating counter-arguments (based on their limited background knowledge and experience of the world) that "trump" the parents' arguments. It would be useful to study such examples.

Echoing Code's work on **epistemic responsibility** and Messick's (1989) work on consequential validity, a number of the contributors to this book argue that the goodness of critical thinking is tied not only to good argument and judgment but also to their consequences in particular contexts. The interest here is not merely in thinking but also in doing. Pinto and Portelli, in the first edition of this volumn, are concerned that the teaching of critical thinking has been tied in recent years to a utilitarian business interest in education — suggesting that critical thinking should be situated within the framework of larger social democratic ideals. It seems likely that all the contributors to this volume would agree that consequences are of great importance. Indeed, since its inception the critical thinking movement has been rooted in the conviction that critical thinking will positively affect our personal, social, and political lives.

One might readily compare the ideals of critical thinking to Code's account of wisdom, which she regards as virtually interchangeable with epistemic responsibility. Wisdom, she asserts, "involves knowing what cognitive ends are worth pursuing and understanding the value of seeing particular cognitive endeavors in context so as to achieve a just estimation of their significance" (53). This emphasis on context and consequences in wise thinking is also evident in Sternberg (2003), who characterizes a wise person[4] as someone with the following qualities:

a. reasoning ability (involving knowledge, logic, reasoning);
b. sagacity (being open to advice, being fair, acknowledging

error, showing concern for others);

c. the ability to learn from ideas and the environment;

d. judgment (understanding self-limitations, undertaking thoughtful action, considering the long as well as the short view of things);

e. the ability to use information expeditiously; and

f. perspicacity (having intuition, reading between the lines, positing solutions on the side of rightness and truth). (178-9)

To act with wisdom is, arguably, the ultimate goal in critical thinking, because it integrates thinking and action within the context of a broader good. But even if there were general agreement about this claim (which there is not), it would still leave open the tasks of determining how critical thinking — broadly or situationally defined — might be assessed, and how the goodness of a particular critical episode might be evaluated.

Epistemologies of Assessment?

Any assessment of critical thinking immediately intersects with explicitly or implicitly held knowledge claims about assessment in general. The expression of these claims in any particular assessment can be considered in terms of four components: (a) the architecture/ technology undergirding the assessment, (b) how the goodness value-scale (the absence, presence, or abundance of knowledge or ability) is codified within the assessment, and (c) the contextual sensitivity of the assessment.

The architecture/ technology of assessment

The architecture or technology[5] of an assessment is like the basic design of a building: it enables certain activities but not others. Assessment design in critical thinking, like assessment design elsewhere, invites and enables the scrutiny of some elements of the area being assessed and does not enable or invite the scrutiny of others (Murphy 2003a). Assessment design seems so fundamental that it is remarkable to find Snow lamenting, as recently as 1993, that "the field of educational and psychological testing suffers today because it never developed a psychology of test design" (45).[6] Mounting criticisms of multiple-choice testing and concern about the broader implications of assessment using *p* constructed responses led Snow (1993) and Bennett (1993) to develop hierarchies of assessment design (see *Table 1*). In these hierarchies, multiple-choice assessment occupies the lowest level, whereas "presentation" or collection of different assessments occupies the highest.

The hierarchies suggested by Snow and Bennett represent emergent possibilities of the relative goodness of assessment design types in a relatively under-theorized and under-investigated area in psychology. Given the lack of development in assessment design as a field,[7] we must not expect any assessment design to afford perfection in assessment. Just as building designers must face up to design trade-offs (because of costs, material availability, aesthetics, zoning regulations, etc.), assessment designers *and* users[8] must face up to the trade-offs and limitations inherent in any single assessment design. As Code (1987) asserts, the goodness of any assessment must be bounded by a recognition of its limitations within specific contexts.

Consider, for example, multiple-choice assessments. Much has been written about the goodness of and the problems associated with standardized multiple-choice assessments (e.g., Ennis, Giancarlo, Groarke, and Sobocan, this volume; Hill and Larsen 2000; Murphy 1994, 2001, 2003b; Murphy, Shannon, Johnston, and Hansen 1998). As is the case for most things, goodness in relation to these assessments can be judged on two levels. On one level, the issue of goodness is about the specific design genre incorporated in an assessment (see, e.g., Johnson, this volume). On the other level, the goodness at issue is whether the specific design genre was well implemented in a particular assessment tool. In the latter case, the question would be whether the tests were well designed in relation to the principles undergirding good multiple-choice test design (see Groarke, this volume; Hill and Larsen 2000; Murphy, Shannon, Johnston, and Hansen 1998).

Bennett's Hierarchy		**Snow's Taxonomy**	
Design Feature	*Design Description*	*Design Feature*	*Design Description*
Multiple choice	Choose from an array of options	Multiple choice	Choose from an array of options
Selection/ identification	Number of choices large enough to eliminate guessing (e.g., key lists)	Multiple choice with intervening construction	Retrieve, reconstruct, reason with knowledge
Reordering/ rearrangement	Place in correct or alternate sequence	Short-answer essay, complex construction	Generate sentence or paragraph
Substitution/ correction	Replace with alternative	Problem exercise	Generate/ explain solution
Completion	Complete sentence, problem with single numerical response, etc.	Teach-back procedure	Explain concept, procedures, structure, system
Construction	Construction of unit such as graph, written explanation, drawing, proof	Long essay, demonstration, or project	Produce with or without topic constraint
Presentation	Physical presentation or performance using real or simulated conditions	Collections of above over time, portfolios, etc.	

Table 1: Proposed categorizations for assessment design. Adapted from Bennett (1993, 3-4) and Snow (1993, 48).

Like multiple-choice assessments, other assessment designs must answer both of these goodness questions: (a) Is the design genre of the assessment a good one? and (b) Is the implementation of this specific assessment design genre well done? Much has been written recently about performance-based or “authentic” assessments that are typically intended as a counterpoint to the shortcomings of standardized multiple-choice tests, or as a way to offer more opportunities for more cognitively demanding responses (see Giancarlo-Gittens, and Sobocan, this volume). Although there may be something to these motivations, even performance-based or “authentic” assessments have some general shortcomings (Snow 1993; Murphy 1995), and are likely to have additional shortcomings when used in relation to specific contexts. It bears repeating that no assessment is perfect.

Goodness in any assessment design demands its own set of warrants which must be considered in relation to contexts of use. Perhaps because of a culture of assessment in the United States (Hanson 1993), and despite a relatively under-theorized design basis, the goodness of multiple-choice standardized assessment has been assumed by society at large. Such assumptions are ethically untenable and, like all assessments, multiple-choice standardized assessments must be judged in relation to specific contexts of use.

The codification of goodness

How an assessment is used in context is inevitably intertwined with how that assessment codifies the goodness factor. In the context of a critical thinking assessment, codification provides an answer to the question of how good an individual’s critical thinking is.

Codification is an attempt to measure the relative degree of knowledge, skills, or ability. The codification of goodness in assessment has two basic facets: an encoding mechanism — that is, whether the summary statements describing the goodness are reported numerically or in words; and a comparison index —

the thing to which any single performance on an assessment is being compared. Assessment design decisions about these two facets introduce assumptions about knowledge into assessment and thereby reveal goodness in particular ways. The different aspects of codification can be represented as in *Table 2* below.

	Encoding Mechanism	
Comparison Index		
Relative to how others perform	Numbers	Words
Relative to descriptions of knowledge, skills, and concepts	Numbers	Words

Table 2: The coding of goodness in assessment

The numerical codification of goodness has been a common feature of schooling for the past century. Percentages, percentile ranks, and standard scores are among the many ways in which numbers are used to summarize a person's (or a school's) goodness on an assessment. Minimalism is both a strength and a weakness of this approach. Numerical codification is designed to distill down to a single number, or a series of numbers, the essence of a performance. Such numerical descriptions are often seen as more efficient than elaborate word-based codifications of assessment performance.

But fundamental questions have been raised about numerical codification. Hacking (1990) argues "that defining new classes of people for the purposes of statistics has consequences for the ways in which we conceive of others and think of our own possibilities and potentialities" (6). The statistical transformations that result play a central role in the "making up of people" (ibid.). In assessment, the numbers encoding the assessment are transformations of performance into some form of countable unit. This numerical distancing creates a veneer of mathematical precision replete with the social values accorded to such precision.

On top of this numerical transformation often comes another transformation whereby numbers are re-transformed into simple categories (e.g., above average, below average, gifted, and so on) which take on a special meaning and contribute to the "making up of people" that Hacking proposes. These categories of re-transformation go on to assume a consequential burden (through the social effects of labeling, access to selected types of education or goods, and so on) that extends far beyond the limited context of the assessment in which they were achieved (Murphy 2003b).

Beyond the categorical dilemmas associated with the numerical codification of goodness is a series of other mathematical issues. The first is one of language. To take but one instance, "percentage" and "percentile" represent quite different aspects of a data set despite the similarity of their names. A second issue arises when one considers what numbers are understood to represent. If a student receives 80 percent on an in-class test, for example, what does this number represent? A grade based on correct responses to weighted or unweighted questions? A grade based on a comparison with the performance of others? One based on the result of past performances of the person being tested? Or one of specific proportion of the knowledge presented in the class? Unless test designers and interpreters can provide reasonable responses to these types of questions, the assumptions underlying the numerical scale used are open to question.

Overarching both of these mathematical issues is the concern that numbers have very specific properties and are based on assumptions that should not be violated. When numbers are used in a categorical sense (as labels in rating scales, with, for example, 1 representing high and 5 representing low), they should not be treated as though they were pure numbers and added, subtracted, divided, and multiplied, because such operations further distort the results of the performance.[9] Mathematical operations result in, for instance, an overall score of average for someone who is rated low on 50 percent of the items and high on the remaining items. Yet, the fallacy of such mathematical manipulations becomes clear when the object of the rating is the bitterness or sweetness of a food. If half the items tasted are rated bitter and half sweet, it would be absurd to say that the food being tasted was neither bitter nor sweet but somewhere in the middle. Yet, these types of inferential leaps based on mathematical operations are routinely made when dealing with ratings involving abstract or difficult-to-pin-down concepts such as critical thinking.

The word-based codification of performance may not have mathematical issues to contend with, but it has issues of its own. First, the word-based encoding of assessment performance runs the gamut from lengthy verbal descriptions of performance to single-word descriptors of performance (poor, average, very good, etc.) that may or may not be based on a set of rubrics. In comparison with numerical codification, word-based codification may be more sensitive to the nuances and context of actual performance simply because numbers are tightly constrained by rules and assumptions whereas words may not be as constrained.

Lengthy descriptive word-based codifications tend to particularize the performance to an immediate context and the meaning of word-based codifications may be more transparent than the meaning of numerically based assessments. But single-word codifications offer a minimalist description which raises many of the same issues found in numerical codification. Such codifications may be based on rubrics or descriptions that are more elaborate word-based codifications of the knowledge, skills, or concepts at stake in the assessment, but the transformation to single-word descriptions raises many of the same issues as numerical codifications. As anyone who reads literature must confess, words offer their own challenges in terms of interpretation — defensibility or warrantability of interpretation is, therefore, a key component of word-based codifications of performance.

Whether numbers or words are used to codify, perform codification is a relative task. The encodings of performance depend performance upon that to which the performance is compared: the performance of other people (including oneself) performing the same set of tasks (e.g., norm-referenced assessment, classroom ranking of student performance); or an ideal description of the concepts and skills within a field of knowledge (e.g., criterion-referenced assessment, items on an in-class test). Any assessment statement must be understood in the context of the interaction between an encoding mechanism and a comparison index, and the possibilities and issues that this interaction raises.

Contextual sensitivity in assessment

For much of the past century, one dominant goal in assessment was the development of assessments that were thought to be impervious to context. Such assessments were designed in a standardized fashion, not unlike quasi-experimental research, whereby as many sources of extraneous influence (variance) as possible were controlled. The idea was that the resultant performance would reflect the residue of pure knowledge/ skills/ abilities that transcend the imagined bounds of specific contexts. This imagined state, in which standardized assessments have no context, allowed comparisons among individuals on a fixed set of tasks,[10] but these regulated standardized assessment contexts created contexts so unique that no analogous contexts existed elsewhere, raising the question of what the assessment showed about performance in non-assessment contexts.

Standardized, typically norm-referenced, assessments can be contrasted with assessments for which context is everything. Performance-based or "authentic" assessments that involve simulation or real-life assessment situations are about the here-and-now. Of course, the devilish question for these assessments is whether they can offer much of relevance beyond the immediate context. In situations such as the high schools of Central Park East, where a portfolio presentation/ defence system is in place (see Darling-Hammond, Ancess, and Falk 1995), the generalizability of skill, knowledge, or ability is often captured because performance is interpreted as representative of the student's learning. The interpreter of the assessment "reads into" the assessment (Moss 1994) instead of having the assessment task "announce" its judgment via the process of numerical transformations. In this way, as Giancarlo-Gittens argues, performance-based or "authentic" assessments are more open ended than standardized multiple-choice tests.

In keeping with current theorizing about assessment and with conceptions of epistemically responsible action, the results of assessment must be interpreted in context. This means that warrantable claims about the critical thinking skills captured in a dormant assessment must work in tandem with warrantable claims about the use of the assessment for specific ends in specific contexts in order for the assessment to be considered valid (Messick 1989). This is an inherently reasonable stance but one need look no further than many governmental policies to see the damaging consequences of assessments that are not properly considered in relation to their contexts (Meier and Wood 2004).

Critical Thinking Assessment:
The Epistemological Double Helix

Again, no assessment is perfect. In the context of critical thinking assessment, further issues are raised by the lack of a definitive description of critical thinking. The challenge that this places on designers and users of critical thinking assessments is to start from these premises, but not to make it impossible to say anything worthwhile about critical thinking. Rather, this double helix of imperfection can be the basis of a commitment to epistemic responsibility in this case — a commitment to consider carefully both the definition of critical thinking and the design of the assessment. Designers and users must be open to the possibility of enlarging, narrowing, refining, elaborating, or discarding definitions and assessments in relation to the particular contexts of use. This double helix of imperfection obligates them to continue to try to perfect concepts and design while simultaneously accepting that perfection is unattainable.

Once the imperfection of the concepts undergirding assessments of critical thinking is recognized, it seems imperative that all assessments must be viewed with a moderating eye in terms of the consequences they may have (for assessors and those who are assessed). To avoid acting in an epistemically irresponsible way, designers and assessors need to be somewhat circumspect; to consider how goodness is and is not instantiated in any one assessment activity; and to demand that multiple assessment artifacts of different types be assembled to warrant any claim of significant consequence. This epistemically and ethically driven conduct, with its obligation to be open and contextual, may in turn offer new routes in the development of both critical thinking and assessment. This book is a concerted and worthwhile effort in the direction of fulfilling such epistemic and ethical obligations in relation to the goodness of critical thinking assessment and evaluation.

References

Belanoff, P., and M. Dickson, eds. 1991. *Portfolios: Process and product*. Portsmouth, NH: Boynton Cook Heinemann.

Bennett, R. 1993. On the meanings of constructed response. In *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*, ed. R. Bennett and W. Ward, pp. 1-27. Hillsdale, NJ: Lawrence Erlbaum.

Bormuth, J. 1970. *On a theory of achievement test items*. Chicago: University of Chicago Press.

Brainy Quote. 2005. Alec Baldwin quotes. Online. Available at http:// www.brainyquote. com/quotes/quotes/ a /alecbaldwil79389. html.

Case, R., and S. Stipp. 2008. Assessment strategies for secondary classrooms. In *The anthology of social studies*. Vol. 2, *Issues and strategies for secondary teachers*, ed. R. Case and P. Clark, 383-97. Vancouver: Pacific Educational Press.

Code, L. 1987. *Epistemic responsibility*. London: University Press of New England.

Critical Thinking Company. 2005. Online. Available at http://vvww.criticalthinking. com/index.jsp.

Darling-Hammond, L., J. Ancess, and B. Falk. 1995. *Authentic assessment in action: Studies of schools and students at work*. New York: Teachers College Press.

Fillmore, C. 1982. Ideal readers and real readers. In *Analyzing discourse: Text and talk*, Georgetown University Roundtable on Languages and Linguistics-1981, ed. D. Tannen, 248-69. Washington, DC: Georgetown University Press.

Fillmore, C., and P. Kay. 1983. *Text semantic analysis of reading comprehension tests*. Berkeley, CA: Institute of Human Learning, University of California (ERIC Document Reproduction Service, ED 238 903).

Fleck, L. 1979. *Genesis and development of a scientific fact*, ed. T. Treun and R. Merton, trans. F. Bradley and T Treun. Chicago: University of Chicago Press.

Goodman, K., E. Smith, R. Meredith, and Y. Goodman. 1987. *Language and thinking in school: A whole language curriculum*, 3d ed. New York: Richard Owen.

Hacking, I. 1990. *The taming of chance*. Cambridge: Cambridge University Press.

Haladyna, T. 1999. *Developing and validating multiple-choice test items*. Mahwah, NJ: Erlbaum.

Hanson, F. 1993. *Testing testing: Social consequences of the examined lift*. Berkeley, CA: University of California Press.

Hill, C., and E. Larsen. 2000. *Children and reading tests*. Vol. 65 of *Advances in discourse processes*. Stamford, CT: Ablex.

Korsgaard, C. 2005. Theories of the good. In *The shorter Routledge encyclopedia of philosophy*, ed. E. Craig, 325. London: Routledge.

Levy, M., and M. Salvadori. 1992. *Why buildings fall down: How structures fail*. New York: Norton.

Meier, D., and G. Wood, eds. 2004. *Many children left behind: How the No Child Left Behind Act is damaging our children and our schools*. Boston: Beacon Press.

Messick, S. 1988. The once and future issues of validity: Assessing the meaning and consequences of measurement. In *Test validity*, ed. H. Wainer and H. Braum, 33-45. Hillsdale, NJ: Erlbaum.

Moss, P. 1994. Can there be validity without reliability? *Educational Researcher* 23(2): 5-12.

Murphy, S. 2003a. Finding literacy: A review of the research on literacy assessment in early childhood education. In *Handbook of early childhood literacy*, ed. N. Hall, J. Larson, and J. Marsh, 369-78. London: Sage.

____. 2003b. Literacy assessment and the politics of identities. In *Contextualising difficulties in literacy development: Exploring politics, culture, ethnicity and ethics*, ed. J. Soler, J. Wearmouth, and G. Reid, 87-101. London: Open University Press.

____. 2001. “No one has ever grown taller as a result of being measured” revisited: More educational measurement lessons for Canadians. In *The erosion of the democracy in education: Critique to possibilities*, ed. J. Portelli and P. Solomon, 145-67. Calgary, AB: Detselig/Temeron Books.

____. 1995. *Revisioning reading assessment: Remembering to learn from the legacy of reading tests*. Clearing House 68: 235-9.

____. 1994. “No one ever grew taller by being measured”: Six educational measurement lessons for Canadians. In *Sociology of education in Canada*, ed. L. Erwin and D. MacLennan, 238-52. Toronto: Copp Clark.

Murphy, S., P. Shannon, P. Johnston, and J. Hansen. 1998. *Fragile evidence: A critique of reading assessment*. Mahwah, NJ: Erlbaum.

Osterlind, S. 1989. *Constructing test items*. Boston: Kluwer.

Snow, R. 1993. Construct validity and constructed response sets. In *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*, ed. R. Bennett and W. Ward, 71-88. Hillsdale, NJ: Erlbaum.

Sternberg, R. 2003. *Wisdom, intelligence and creativity synthesized*. Cambridge, MA: Cambridge University Press.

Underwood, T. 1999. *The portfolio project: A study of assessment, instruction, and middle school reform*. Urbana, IL: National Council of Teachers of English.

Notes

1. When arguing for a knowledge claim, one is often arguing within an epistemic community - a community that shares in large part a similar knowledge base. In such a situation, a community's interest in consensus among its members (which represents a type of good - in the minds of community members, at least) may be taken to override the goodness of epistemic claims even when those claims are justified. One thinks of Galileo-like situations in which a lonely scientist who has made an innovative discovery is confronted by a community that

refuses to consider what are much later considered to be well-warranted claims because they would shatter the grounds upon which that community is founded. In these and other cases, epistemic communities continually struggle with the tension between convention, which preserves communal knowledge, and invention, which challenges such knowledge but has the potential to move a community forward in its collective thinking.

2. Understood as a set of ideas or propositions that does not necessarily include a contestational stance.

3. This position fails to consider its inherent temporal paradox - simply put, how did the "first" critical thinker acquire his or her abilities?

4. Sternberg (2003) offers a theory that relates wisdom, intelligence, and creativity. The listing of some of the traits of wisdom does not do justice to the complexity of his theory but is offered here as a possible direction for the elaboration of theories of critical thinking.

5. Technology is used here in the same sense that a technology invites particular uses. For instance, a pencil invites use as a tool for writing. Granted, the pencil may be used for many other things, depending on context, but the principal invitation extended by the tool is to write.

6. For instance, much emphasis has been placed on statistical aspects of multiple-choice tests, but only a handful of texts have been written on the design of multiple-choice test items despite the fact that they have been heavily used for much of the past century. Bormuth (1970) writes about controlling confounding sources of error in tests through highly controlling the language in the tests while Haladyna (1999) and Osterlind (1989) both focus on a range of multiple-choice test design issues.

7. Although the overall field of assessment design is underdeveloped in psychology, it should be noted that some individual assessment types have been the focus of much attention. For instance, with the adoption of portfolio assessment within the field of education, numerous texts have been devoted to considering what the components of portfolio assessment ought to be and how portfolio assessment should be implemented (e.g., Belanoff and Dickson 1991; Underwood 1999; Case 2008).

8. Designs are plans for anticipated uses. However, human beings, being social and inventive, often use things for purposes other than those anticipated by designers. Many examples of the unanticipated uses and

stresses on buildings are documented in the text *Why Buildings Fall Down: How Structures Fail* (Levy and Salvadori 1992). For instance, an atrium bridge in a hotel lobby may well have met design specifications for large crowds but when the crowds unanticipatedly jump up and down to the beat of music, new and unanticipated stresses occur. In building design, the goal is to anticipate more and more uses and design assessments that take into account those uses. In assessment, designers place the burden of use on the users. For instance, in most large-scale standardized test manuals there are warnings about overreaching the meaning of the findings (see Murphy et al. 1998). Indeed, in individualized assessments, the same is true (ibid.). However, a look at press headlines, political statements, or the statements of many others (Murphy 1995, 2001; Pinto and Portelli, 1st ed. of this volume) indicates that assessment results can be used well beyond the purposes for which they were intended (see Ennis, this volume, for a discussion of purposes of assessment). The consequential validity (Messick 1988) of these assessments is thus put in question. Unfortunately, however, some users treat assessment results as definitive rather than as the tentative and fragile (Murphy et al. 1998) pieces of documentation that they are.

9. A good example of the inappropriateness of mathematical operations for the categorical use of numbers can be illustrated by an example relating to taste, where 1 is very sweet and 5 is very tart. If we have two very tart drinks and two very sweet drinks, would it be appropriate to say that on average the drinks tasted medium — not too sweet and not too tart? Most people would agree that such labeling does not represent the tastes but would agree that a better representation would be to say that two drinks are very tart and two very sweet. Yet when the 1 is assigned the label "very good" and the 5 is assigned the label of "poor," many fail to see the fallacy in saying that the performance yielded from two scores of 5 and two scores of 1 is average.

10. Of course, there is also a fundamental assumption at work in standardized assessment design — the assessments have been very well designed. Yet, a variety of sources (e.g., Fillmore 1982; Filmore and Kay 1983; Murphy et al. 1998; Hill and Larsen 2000) suggest that fairly recent examples of such assessments reveal relatively poor design. Added to the poor design features are problems of use. For instance, there may be unmet assumptions that the characteristics of persons taking the assessment are similar to the characteristics of those

upon whom the test was originally normed.