University Press of Colorado
Utah State University Press

Chapter Title: ETHICS, PSYCHOMETRICS, AND WRITING ASSESSMENT: A Conceptual Model
Chapter Author(s): Robert J. Mislevy and Norbert Elliot

Book Title: After Plato
Book Subtitle: Rhetoric, Ethics, and the Teaching of Writing
Book Editor(s): JOHN DUFFY, LOIS AGNEW
Published by: University Press of Colorado, Utah State University Press. (2020)
Stable URL: https://www.jstor.org/stable/j.ctv13qfw0s.11

SECTION TWO

Disciplinary and Pedagogical Perspectives

8

ETHICS, PSYCHOMETRICS, AND WRITING ASSESSMENT
A Conceptual Model

Robert J. Mislevy
Norbert Elliot

Providing a coherent conceptual model for relationships among ethics, psychometrics, and writing assessment is a complex undertaking. Challenges are apparent at a moment's thought: the landscape is vast, sources of evidence are field-dependent, and the points of entry are contested. A possible way forward is to propose a broad definition of ethics and then to narrow that definition as its elements apply to writing assessment. Psychometrics may then be added to the model as a sociocognitive framework in which language is understood as situated and constitutive. In this way, we may then proceed to the details of a conceptual model for writing assessment informed by the fields of ethics and psychometrics.

Let's begin with a basic definition of ethical philosophy proposed by James Rachaels (1986, 11)—his minimum conception of morality:

> Morality is, at the very least, the effort to guide one's conduct by reason—that is, to do what there are the best reasons for doing—while giving equal weight to the interests of each individual who will be affected by one's conduct.

Rachaels (1986, 11) extends the definition by linking it to the capabilities needed to be a conscious moral agent:

> Someone who is concerned impartially with the interests of everyone affected by what he or she does; who carefully sifts facts and examines their implications; who accepts principles of conduct only after scrutinizing them to make sure they are sound; who is willing to "listen to reason" even when it means that his or her earlier convictions may have to be revised; and who, finally, is willing to act on the results of this deliberation.

This definition has distinct advantages as the general philosophical background for our conceptual model of writing assessment. Published in

1986, the definition benefits from three philosophical traditions—the categorical imperative of Immanuel Kant (2012 [1785], 31): "Act only in accordance with that maxim through which you can at the same time will that it become a universal law"; the greatest happiness principle of John Stuart Mill (1879 [1861], 17): "The ultimate end, with reference to and for the sake of which all other things are desirable (whether we are considering our own good or that of other people), is an existence exempt as far as possible from pain, and as rich as possible in enjoyments, both in point of quantity and quality"; and the social justice theory of John Rawls (1999 [1971], 96): "We are not to gain from the cooperative labor of others without doing our fair share." Informed by philosophical perspectives of conscience, capability, and community, it seems wise to adopt a complementarity model yielding overlapping consensus an endorsement of principles, as philosopher Martha C. Nussbaum (2011) observes, for a given purpose and not as a guide to life. Put straightforwardly, we advance complementarity in our conceptual model as an overarching way to establish new directions for research leading to the advancement of individual capability.

Extending this ethical framework—a sense of reason tempered by consequence understood as particularized, convictions revisited by reflection, and fairness enacted in communities—we now offer a definition of ethics in writing assessment:

> Fairness in writing assessment is defined as the identification of opportunity structures created through maximum construct representation. Constraint of the writing construct is to be tolerated only to the extent to which benefits are realized for the least advantaged. (Elliot 2016, §3.1)

In advancing a minimum concept of morality that informs the actions of a conscious moral agent, this definition provides needed focus. First, it yields an overarching frame in which the advancement of opportunity is understood as the aim of writing assessment. Here we recall James Paul Gee (2008) and his identification of the rights of students in terms of opportunity to learn: universal affordances for action, participation, and learning; attention to experiential ranges of students; equal access to relevant technologies; accountability for the communities of practice who manage that information; and emphasis on identity, value, content, and characteristic activities associated with language across academic areas. Thus the definition's emphasis on opportunity to learn plays a unique role. Second, the definition aims to lessen the social inequality resulting from writing assessment while emphasizing the benefits to be gained by assessment. Third, the definition bridges the philosophical

and the psychometric with its emphasis on fairness in relationship to validity and reliability. Here the definition leverages the 2014 edition of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014) in which fairness in testing has been considered as its own category of evidence alongside foundational concepts of validity and reliability. Fourth, the definition advances construct-based assessments—those assessments that have at their core a well-defined, evidence-based conceptualization of the writing construct—and links those assessments to fairness. Thus the definition implicitly advances recent efforts by the Institute of Education Science to bring forward recommendations of writing instruction based on empirical evidence meeting exacting research demands (Graham et al. 2012, 2016). Since the definition takes as the aim of writing assessment advancement of opportunity to learn, teaching and assessing writing are taken as inextricably interrelated under defined construct models. Fifth, the focus of the definition on opportunity structures aligns the Western philosophical tradition with non-Western traditions. As Nussbaum reminds us, emphasis on human rights and capabilities exists in both Indian and Chinese traditions. Thus the philosophers we reference, although Western, need not lead us to value imperialism. As Bo Wang argues in this volume, Confucian perspectives provide alternatives to the values of autonomy and individualism. Our definition of ethics in writing assessment is therefore open to all philosophical traditions that value the pursuit of human capabilities. With Rawls and Nussbaum, we affirm those philosophical traditions in which diverse perspectives, formed through collaboration, are used to advance justice and capability.

The standpoint we are proposing in this chapter is therefore a confluence of three lines of work: the growing understanding of fairness, validity, and reliability as both categories of evidence and ethical frameworks (Kelly-Riley and Whithaus 2016); an ecological view of human cognition as applied to writing (White, Elliot, and Peckham 2015); and a sociocognitive view of assessment as an expression of contextualized, purposive arguments (Mislevy 2018). Significantly, our model thus resonates with that of David Slomp's (2016) framework for using consequential validity evidence in evaluating each step in writing assessment design. These lines of work inform the structure of our chapter and the model proposed in it.

We first turn to an integrative perspective in which a defined view of psychometrics (identified as sociocognitive) is used to design a

construct-driven approach to writing (identified as nomothetic). We then combine the desire to use situative perspectives with efforts to describe the span of the writing construct. We conclude with a summary of four ethical principles that are implicit in our model that, taken together, call for enhanced attention to construct validity and the integrative framework that is provided by evidence of fairness. In emphasizing this integrative approach, we believe researchers can structurally design assessments that, in context and use, are ethical in nature. With the reservation that conceptual models are limited in terms of their applications—as are all human endeavors intending to structure opportunities—we hold that ours bakes a certain kind of ethical perspective (identified with fairness) into the assessment itself.

Our proposed model comes with a warning label for our readers. Our work is based on two fields of research—psychometrics and writing assessment—in which empirical techniques are used to gather evidence of fairness, reliability, and validity. Our chapter is therefore complex at times in its use of terminology. As well, our discussion is based on current empirical knowledge of educational measurement in general and writing assessment in particular. Ellen Schendel and Peggy O'Neill (1999) were among the first assessment scholars to remind us that quantitative research is an ethical obligation to our students, and we extend their belief here—with eyes wide open to the costs and effectiveness when mathematical models are used to make inferences about human behavior. The decisions we make when we prepare our writing assessment programs nevertheless raise questions that can only be answered empirically: Are there performance differences among sub-groups that are related to the writing assessment task? Does the assessment task operationalize an explicit model of the writing construct that all groups may access equally? Is there inter-topic reliability among those tasks for all student groups? Required, therefore, is a willingness to engage current programs of research in fields that are uncommonly intricate.

Our model also comes with an admission. The chapter seeks to extend an existing theory of ethics for writing assessment by emphasizing the benefits to be gained from psychometrics—key concepts and methods, that is, as opposed to rote applications of long-standing practices. We therefore seek a restorative vision in which quantitative methods are seen as one important way of understanding students as they make language visible on notebook pages and computer screens. Many of our colleagues continue to depict assessment as a uniform activity of oppression, and their work is important if we are to understand the consequences of our actions. Our vision, however, carries nothing

of the tacit conservatism inaccurately attributed to anyone who uses traditional and advanced empirical techniques in educational research. As such, we invite our readers to understand our work as driven by a liberal imagination.

PSYCHOMETRIC MODELING: A SOCIOCOGNITIVE PERSPECTIVE ON ASSESSMENT

Paul Holland, a former president of the Psychometric Society, described the history of psychometrics in educational and psychological assessment, or test theory, in terms of four generations (Holland 2008). Neil J. Dorans (2011, 259) summarized the first three like this:

> The first generation, which was influenced by concepts such as error of measurement and correlation that were developed in other fields, focused on test scores and saw developments in the areas of reliability, classical test theory, generalizability theory, and validity. This generation began in the early twentieth century and continues today, but most of its major developments were achieved by 1970. The second generation, which focused on models for item level data, began in the 1940s and peaked in the 1970s but continues into the present as well. The third generation started in the 1970s and continues into today. It is characterized by the application of statistical ideas and sophisticated computational methods to item level models, as well as models of sets of items.

Most large-scale assessments in writing and other disciplines remain grounded in first-generation concepts and methods even while drawing on second- and third-generation machinery to improve quality and efficiency to the same ostensive end: to measure a trait that is qualitatively the same for all the test takers, who can be assessed with samples of the same tasks, have their performances evaluated in the same way, and have their results reported as locations on the same proficiency scale. In the second generation, the mid-century emergence of Item Response Theory (IRT) under psychometricians such as Frederic M. Lord (1952) allowed for the more flexible designs in the National Assessment for Educational Progress. Generalizability theory, developed by Lee J. Cronbach and his colleagues (1972), enabled illuminating analyses of ratings of multiple aspects of students' performances and the consistency of evaluations of multiple raters. Third-generation efforts draw on elements from IRT and generalizability and other techniques such as latent class analysis, hierarchical modeling, diagnostic classification, and mixture models to build models tailored to assessment designs, response formats, aspects of performance, and student-population structures (see, for example,

De Boeck and Wilson 2004). Third-generation psychometrics assimilated emerging technologies, with computer-based testing and automated scoring of essays to produce scores to match humans' ratings.

Not much of this history of test theory matches our narratives of interactions among writing teachers and their students. Measurement error seems to have little bearing on, for example, the ways teachers help student prepare initial drafts for an assignment. These interactions address no single writing proficiency. Rather, the interactions encompass unique aspects of writing as drawn down in the writing task at hand. Classroom and online discussions focus on the particulars of genre, purpose, audience, context, and the language evoked by the task. Ever situative, students have different backgrounds, goals, and writing experiences to build on as they engage the task. Teachers do not merely take into account such contexts; rather, using process-based approaches, teachers build on the situatedness of each student in ways that foster individual learning. Teaching writing comes down to finding the sweet spot between informed pedagogy for the class and the fit of that pedagogy to the individual student.

One frontier in psychometrics looks in just these directions. Dorans (2011, 259) continues:

> The current fourth generation [of test theory] attempts to bridge the gap between the statistician/psychometrician role and the role of other components of the testing enterprise. It recognizes that testing occurs within a larger complex system and that measurement needs to occur within this larger context.

Here is where writing pedagogy intersects with the measurement theory. We find in the fourth generation growing attention to the sociocognitive foundations for peoples' capabilities, how they develop, and how they are used in a social and cultural milieu—and just how ideas and methods from test theory might be reconceived, extended, and brought to bear in ways that are consistent with what we are learning about learning (Chalhoub-Deville 2003; Mislevy 2018; Zumbo 2007). Figure 8.1, which we will parse shortly, reflects this view of psychometric modeling. In the field of psychometrics, figure 8.1 is a big leap.

Quietly, at the edges of practice, the ground is shifting. The idea is to give attention to modeling patterns in the dynamic interplay of individuals and the social structures—but with models that reflect patterns as we can now view them through a sociocognitive perspective. Valuing the impact of flux, a sociocognitive view sees human activity as a complex adaptive system involving interactions among and within individuals (Gee 1992; Sperber 1996). Although every social situation is unique,
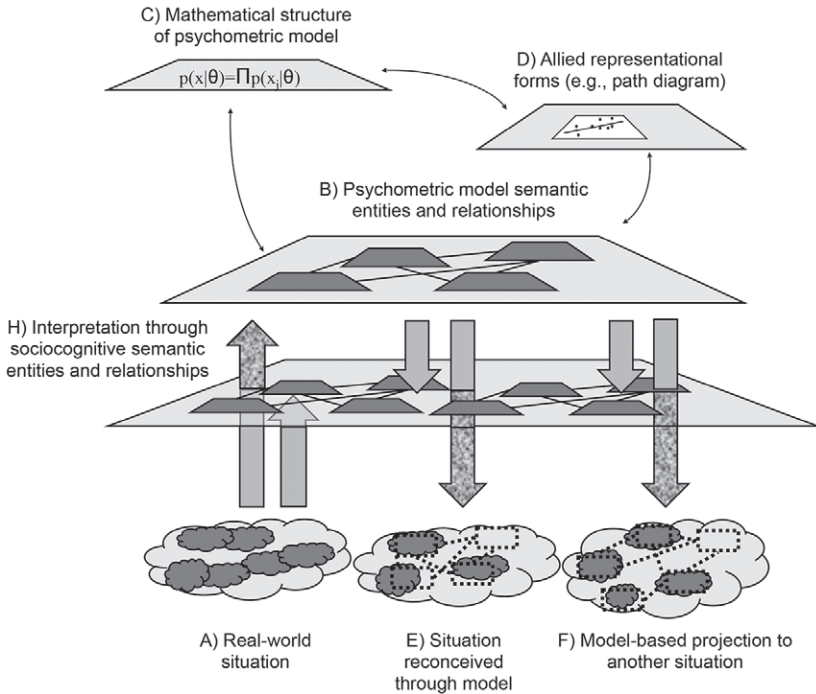
Figure 8.1. Model-based reasoning from a psychometric perspective (Mislevy 2018, reprinted with permission)

regularities emerge in interactions among persons, which collectively might be called linguistic, cultural, and substantive patterns (LCS patterns). It is LCS patterns that make interactions among individuals possible and to which individuals become attuned through experiencing situations constructed around them. Linguistic patterns, for example, span lexicon and syntax; there are typical uses of given words and structures, but every situated meaning depends on context, users' intentions and hearers' expectations, and the interpersonal functions they serve. Genres are examples of cultural patterns—encompassing typical ways people structure writing but intertwined with kinds of purposes and uses people have in recurring kinds of social situations. Substantive patterns address knowledge structures and activity structures in the social and physical world, from repairing toasters to greeting friends to writing research proposals. In every situation, we perceive, act, and interact by assembling resources attuned to LCS patterns of many kinds and at different levels and blending them with the situation at hand (and, in the process, further developing our resources for future experiences).

Figure 8.1 illustrates core properties of model-based reasoning in psychometrics, reconceived through a sociocognitive perspective. It represents aspects of a particular quantitative model as applied in a particular assessment application. At the top left of the figure, representational form C is the mathematical expression of the model, a set of statements in the symbol system of mathematical reasoning and probabilistic expression. Let's take the example of IRT—second-generation models known for their ability to characterize individual item difficulties and test-taker abilities separately but on the same scale, a feature not possible with first-generation models (Lord 1952). IRT contains conditional probability distributions for some variables Xj given another variable θ—which by themselves say nothing about people or performances. Here is the world of mathematical modeling, elegant equations that, if we are fortunate, allow us to learn something about how we are with each other. Representational form D suggests one of perhaps several aspects of the model, such as visualizations, geometric expression of relationships, or computer code for carrying out calculations (as examples, displays of relationships between scores as scatterplots, algorithms for computing Pearson correlation coefficients, and interpreting correlations as cosines between score vectors).

The plane labeled B in the center of the figure, holding a place for psychometric model entities and relationships, is the semantic connection between the equations and the world; it lies in the realm of "as if" thinking (as does all our use of language). There we begin to give semantic expression to concepts involved in the construct at issue, which will take situated meanings in the application. Returning to representational form C, we can now better understand the Xj variables associated with aspects of performance in tasks indexed by j and by θ as an aspect of a person's capability. While these expressions surely do not capture the entirety of the term at hand, the process of establishing a relationship among a concept, the word used to define that concept, and a quantitative expression of some aspects of a real-world situation provides a practical tool for working through certain kinds of complex problems. As a process of enactment, plane B is a gateway to further reasoning.

In real-world situation A at the lower left, the cloud represents the candidates (in our case, the students) sitting for the assessment and their performances. It is here that the concepts used to model entities and relationships come into force. In reconceived situation E, the cloud represents the way we approximate aspects of how that student experiences the assessment within the identified setting in terms of the entities and relationships of the model space represented in plane B.

In the bottom right of the figure, the final cloud, shown in F, represents this assessment as it is envisioned and then extended to another environment—with special attention to the things the student might do in an unknown situation (expressed by that empty box).

Put in motion, figure 8.1 represents a series of considerations and reconsiderations as we move through the model: mathematical reasoning and probabilistic expressions are either expanded or narrowed as they are considered in terms of the model entities and relationships—as are the ways these expressions are visualized (and thus manifested for assessment stakeholders). The real-world situation, in turn, allows reconceptualization of the entities and relationships that are, in turn, further modified as the assessment situation is reconceived and what is learned there is projected (with caveats and qualifiers) to other assessments, instructional activities, or real-world situations.

What is new is the plane labeled H, a reconception of all the variables in the model, all the patterns among aspects of peoples' performances approximated in the model, and all the decisions and inferences supported by the model through a sociocognitive perspective. Importantly, the patterns emerge from their personal histories and the sociocultural context and could be quite different for other test takers and in other sociocultural contexts; careful checking of inferences through the model requires investigation when certain individuals or groups understand tasks or perform qualitatively differently than other students. Reasoning through a model can strengthen inference about individuals by leveraging broader patterns, but it can in these cases distort inferences for certain individuals. The statistical machinery comes with tools to highlight such instances, and a score user has an ethical imperative to use them.

The idea is not to give the same writing tasks over and over and watch some students succeed and others fail. The idea is to advance opportunities to learn for all students by recognizing and making use of the information in the variation. Advances in computer technology and accompanying assessment platforms have enabled modelers to make more nuanced distinctions among student responses to varied tasks. These advancements are not the same, of course, as a perceptive teacher, but even with their limitations, the fine-grained information provided by technology is far more helpful as feedback at a larger scale than an overall measure can be: providing more refined, variegated automated feedback on genre-tuned aspects of a writing sample, for example, or supporting dialogue among peers critiquing each other's work in an internet community, or providing analyses of patterns of performance

across teachers who share assignments and use an online tool to facilitate their evaluations. As a result, students can learn in real time what they know and what they need to know. While, for example, a student may have strong skills in knowledge of conventions, that same student may have difficulty in achieving cohesion in varied writing genres. With this knowledge, instructors can plan in greater detail how to help individual students. The takeaway is straightforward: figure 8.1 is ideally suited to support formative assessment and is, in fact, readily adaptable for settings in which sociocognitive modeling in local settings is desired.

The match between the situation and the model is never perfect, nor do we claim it to be so. Nevertheless, our understanding of the situation—and our ability to make inferences beyond the situation—are now both enhanced by a framework the situation does not in itself possess. We now re-understand (and incompletely understand and sometimes misunderstand) the situation in terms of the patterns the model can provide, built around both the regularities and the variations of students, tasks, and performances.

This may sound good in the abstract, but how do we actually go about building models that are suited to contexts and purposes? We draw on what we know globally from sociocognitive research and studies of the domain at issue and what we know locally about the students, purposes, contexts, and constraints at play for the job at hand. We turn now to an example that uses a representation of the writing construct to help an assessment developer think through the design process.

WRITING ASSESSMENT: A CONSTRUCT REPRESENTATION PERSPECTIVE

As is the case with psychometrics, quietly, at the edges of practice, the ground is also shifting for writing assessment. Research at the present time emphasizes the need for greater construct validity in our research practices (Sparks, Song, Brantley, and Lou 2014) and the turn to social justice as a stance framing the validity and the reliability/precision of our theory building and empirical research (Poe, Inoue, and Elliot 2018). Informed by recent developments such as these, a new view of the writing construct has emerged that is congruent with a more sociocognitive view of psychometric model-based reasoning. Figure 8.2 is a visualization of that construct.

The top row spans the three environments in which students communicate: digital, print, and blurred. The second row captures a language arts framework for writing, including writing, reading, speaking,
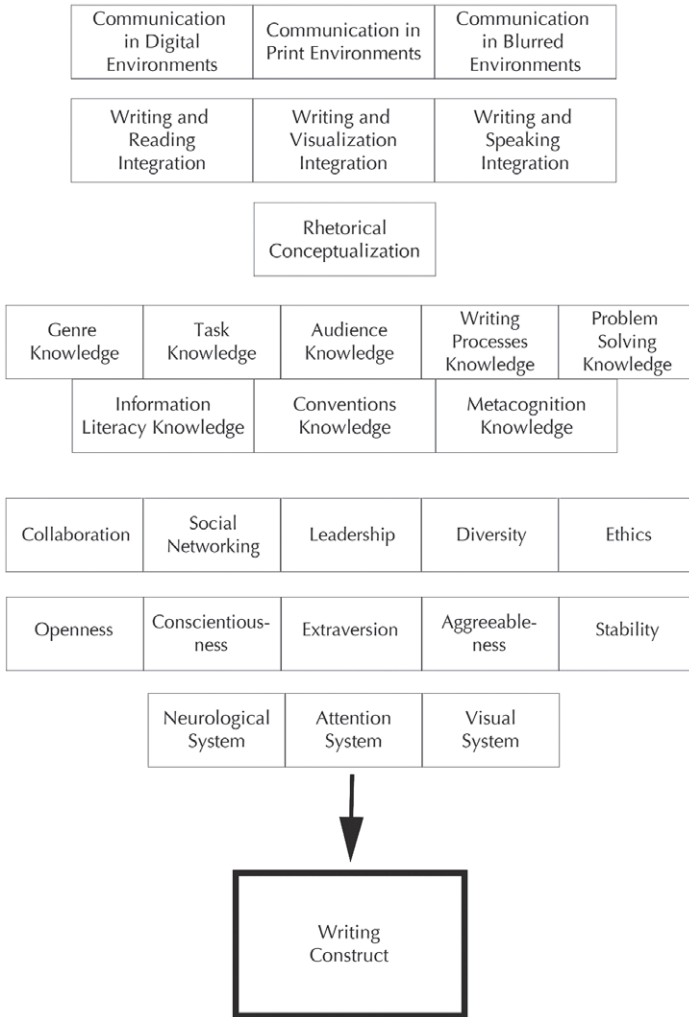
Figure 8.2. Construct representation from a writing assessment perspective (White, Elliot, and Peckham 2015, reprinted with permission)

and listening. The third row identifies rhetorical conceptualization in its attention to language and its sources of knowledge. The fourth and fifth rows identify the cognitive domain of writing. The sixth row identifies the interpersonal domain of writing. The seventh row identifies the intrapersonal domain of writing. The eighth row, often neglected in construct mapping, attends to the neurological, attention, and vision capacities necessary to perform language arts tasks.

As noted above, this view of the writing construct is described as nomothetic—a taxonomy of the span of written communication that is understood to be both constant and general and therefore a way to generate ideographic representations that are detailed and specific. In other words, the rows in figure 8.1 are meant to represent all possible facets of writing as a part of language arts (writing, reading, speaking, and listening). Under this broad taxonomy, specific representations may be created, such as those consensus frameworks used in the *Framework for Success in Postsecondary Writing* (Council of Writing Program Administrators, National Council of Teachers of English, and National Writing Project 2011) and the empirical models derived by John Hayes (2012). The more research we have, the more we can refine the general taxonomy and the better we can create specific models for specific episodes of instruction and assessment. In the field of writing assessment, figure 8.2 is a big leap.

Both figures 8.1 and 8.2 are high-level abstractions. Let's see how they inform and particularize each other when we design a writing assessment in a particular context, for particular students, for a particular purpose. To position the writing construct, imagine sliding figure 8.2 into plane B in the center of figure 8.1 so we can begin the "as if" thinking process. It is here that we begin to enact the semantic expressions by which we give precision to particular communication environments, language arts frameworks, rhetorical conceptualizations, and the cognitive, interpersonal, intrapersonal, and neurological domains of writing that are at issue in the application at hand. While we acknowledge that any definition we give these entities and their relationships is limited, we now have a conceptual gateway to work through our assessment. In figure 8.1, we illustrate how elements of the writing construct interact when they are specified in a given situation—the relationships among the four domains, what we know about the students, what we (and they) need to find out, and what genres and environments are in play. This kind of consideration tells us the kinds of narratives we need to build to support claims about students in the semantic space of the model (B).

In terms of representational form A in figure 8.1, we now begin to consider and select the forms of mathematical reasoning and probabilistic expression—classical test theory, IRT, multidimensional scaling, or something else—that will be of use to situate the writing construct and the psychometric model jointly. That is, we now specify what variables and relationships we need to express in the particularized semantic needed to suit the assessment aim and context. As they are useful to various stakeholders, we also begin to consider the visualizations of these models in representational form B—as well as interpretation (H),

contextualization (A), reconceptualization (E), and generalization (F) based on the construct model. As we turn to the real-world situation, we imagine the individual student and those students of similar and dissimilar backgrounds who will be involved. As we project our thought experiment to other settings, we begin to understand the complexities of the particularized model in terms of interactions within the broader construct model itself as determined by contexts. The model is simplified, but if we have done our work well, it is simplified in a way that is useful for the job at hand. In other work complementary to that which we are presenting in this chapter, we have demonstrated that such particularization may prove useful in supporting the success of diverse student groups in international higher education contexts by improving connections between assessments and instruction for admitted students (Oliveri, Mislevy, and Elliot 2020).

As an example of just one of the many considerations involved in this process, we will do well to reflect on the significance of *genre* in this kind of construct modeling. Carolyn R. Miller (1984, 151) noted that traditionally, "rhetorical genres have been defined by similarities in strategies or forms in the discourses, by similarities in audience, by similarities in modes of thinking, and by similarities in rhetorical situations." She was after something entirely different:

> The classification [of genre] I am advocating is, in effect, ethnomethodological: it seeks to explicate the knowledge that practice creates. This approach insists that the "de facto" genres, the types we have names for in everyday language, tell us something theoretically important about discourse. To consider as potential genres such homely discourse as the letter of recommendation, the user manual, the progress report, the ransom note, the lecture, and the white paper, as well as the eulogy, the apologia, the inaugural, the public proceeding, and the sermon, is not to trivialize the study of genres; it is to take seriously the rhetoric in which we are immersed and the situations in which we find ourselves. (Miller 1984, 155)

In other words, genre is not simply about forms of writing, reading, or speaking but about those forms as ways of using language in recurring situations. Speakers and writers know the kinds of things that are meant to happen and structures their hearers and readers expect to help them produce and to make sense of new information. Becoming proficient means not simply knowing structures but understanding their role in practice, using them to create and to apprehend information in the ways people communicate in those practices. Yes, genre "embraces both form and content . . . [but also] the use of genres simultaneously constitutes and reproduces social structures; and . . . genre conventions signal a

discourse community's norms, epistemology, ideology, and social ontology" (Berkenkotter and Huckin 1993, 475). Writing a research proposal or completing a tax form are genre uses that serve to structure individuals' thinking and acting in a complex network of coordinated activity.

To become proficient in a genre, then, is not simply knowing how to write decontextualized text in a certain form but to write in forms and ways that connect intimately to what people know, what people expect, and what people do in some activity. As John Duffy demonstrates in this volume, alertness to genre theory itself promotes virtue by allowing the formation of bridges (our term is crosswalks) that allow students to understand writing as an achievable task—a form of conceptualization that is important to assessment as we seek to see students at their best. It does no good to simply throw one five-paragraph essay topic after another and watch a student succeed or fail when the goal is to help her improve her skill in crafting résumés or assembling engineering proposals. If we want our students to learn to write in a network of coordinated activities, our curricula cannot be one essay after another. If our aim is to advance opportunity to learn through fairness, then genre selection becomes an ethical matter.

Remaining alert to the power of genre goes a long way in helping us understand topics such as transfer, so important to education in general (National Research Council 2012) and writing research in particular (Yancey, Robertson, and Taczak 2014). A moment's reflection on how genre shifts the writing process makes us realize again the contingent nature of assessment. In their examination of how a professional communication designer constructs a proposal, Mariëlle Leijten and colleagues (2014) acknowledge that the recorded activities—the construction of visual content and the accompanying attention and motivation management—are not apparent in current models of writing. With a genre shift from academy writing (the source-based essay) to professional writing (the proposal), the landscape shifts radically. Without recognizing the impact of such shifts at the lexical and grammatical level, tracking writing transfer becomes impossible.

In the design process, then, we must examine the role of a particular genre in a given assessment, what the students know (and do not know) about the social milieu in which it takes meaning, and what types of feedback will be needed. The genre we highlight in a writing assessment for engineering students, the practices it draws on, and the qualities in performance we attend to will be quite different from those we must address when helping job-seeking adults prepare their résumés—yet both are derived from the same broad writing construct and the same

broad sociocognitive conception of assessment. These considerations shape our decisions about task types, substantive content, evaluation procedures, and the nature and grain size of reports. Together, such contextualizations, articulations, and design choices give the variables and relationships in our models a meaning that is situated and hence useful in the situation.

More generally, educational experiences in school, work, and recreation are meant to help people develop resources associated with targeted LCS patterns and practices that draw on them. Educational assessments are meant to evoke aspects of peoples' diverse capabilities so construed, for a variety of purposes and in a variety of contexts that concern past, present, or potential development. The challenge for assessment and measurement is reconceiving its very foundations accordingly (Mislevy 2018). Models and concepts developed throughout the previous generations can continue to be useful, although with interpretations and uses reconsidered. Others will need to be extended, such as models that characterize the more qualitative aspects of writing performances that are the basis of individual feedback or models that examine patterns of variation and regularity across distributed networks of students and teachers as they critique and support each others' development. Still others will be developed anew, such as incorporating natural language processing tools to supplement learning locally and developing ways to evaluate distributed performances both economically and authentically.

In terms of the offered psychometric perspective and writing construct, their validity is not just the correspondence between the model and the system being modeled. Viewing models as tools people use for reasoning, we recognize that validity depends on the aptness of the relationship among the situation, the model, a user, and a purpose (Suárez 2004). Becoming proficient in the use of models therefore means more than becoming fluent with their mechanics. Becoming proficient means being able to build models to suit real-world situations, evaluating where they fit and when they need to be revised or abandoned, and continually evaluating the quality of reasoning about the real world through the lens of the model. It means understanding which inferences derived through a model are warranted and which are not—as well as which are robust and sensitive to various ways the real-world situation might differ from the model. Psychometric models and assessments are constructive to the extent that they are at once faithful to the construct as we understand it, support sound inferences for their intended purposes, and suit the contexts and constraints of the application at hand.

ETHICAL UNDERSTANDING THROUGH CONCEPTUAL MODELING

Far from a monolith, writing assessment exists along a continuum. At one end, we find standard assessments designed to be distributed across contexts; on the other, we find local assessments intended to provide information tuned to particular students, contexts, purposes, and surrounding information. Along this continuum, the conceptual model we offer, in which model-based reasoning from a psychometric perspective enriches our understanding of the LCS span of the writing construct, will be of varied use. Yet wherever one is located on this continuum, the kind of model-based thought experiment offered in this chapter will be of use in understanding the complexity of the assessment task at hand.

It is within this very complexity that we find the need for ethical understanding. Designing writing assessments and using psychometric models is not a matter of objectively constructing a scientific measure of some unitary capability. It is rather the fashioning of a means to gather, represent, and use information about the resources students can bring in some space of writing activities, for some purpose, in some context, under some constraints. This we can do well or poorly. Issues of evidence and inference arise, and psychometrics can help with those. But equally important are issues of fairness, of opportunity, and of consequences, in various ways to various stakeholders. These are not simply measurement issues but ethical issues, which statistical reasoning may help us clarify but cannot by itself resolve. While we do not have the space here to devote to all the implications of our model, we can conclude with a summary of ethical principles implicit in our approach.

- *There exists an inextricable bond between the technical and the ethical.* This Stoic belief that nothing is good or bad but by its use does not pass the smirk test. With Langdon Winner (1980), we all learned in the twentieth century to realize that artifacts have politics. Harder to recognize, however, is that our technical artifacts are deeply related to our ethical perceptions. The formulas in figure 8.1, representational form A, are based on premises that, in turn, reflect worldviews. Because equations have politics, ethical stance exists at the level of the equation. Without an understanding of the four generations of psychometric research defined above—accompanied by the ability to work with specialists to ensure that attendant measures are used in principled ways—there is as much chance that the writing assessment specialist will throw darkness as cast light. To those who would dismiss empirical knowledge outright, this is a very hard message, but its truth will not go away. If a researcher is involved in the assessment of writing abilities across time and space—and over many students and many instructors—little can be done to help students without an understanding of quantitative methods. Because we need

commonalities between the technical and the ethical, we can do no better than remind ourselves, as does James E. Porter in this volume, that when combined with *techne, praxis* means making something beautiful, positive, and communal. Just as equations have politics, they may also be elegant in their creation of change and their ability to disrupt traditional thinking.

- *There is value in infusing contingency into language-based conceptual systems.* In the interplay between figure 8.1 and figure 8.2, stakeholders participate in a principled system of planning before the assessment episode is begun. In this kind of anticipatory thought, the assessment design runs along an iterative path so that the standpoints of those involved in the assessment community are understood from the beginning. In placing the conceptual system in the referential frame of the test taker, specific attention is oriented to the candidate and others like and unlike that candidate. Here we find useful William Duffy's interpretation of *phronesis* in this volume as a stance requiring balance between what is good and what is expedient, thus encouraging us to rise to the demands of exigencies even as we refuse to foreclose the possibility of future exchange. While seemingly at odds, acknowledgement of contingency and the desire for practical action need not forestall each other. There is an important connection between practical action and anticipatory thought that begins with acknowledgment of contingency.

- *The complexity of a language-based conceptual system should reflect both the intracity of human capability and the humility required to engage that capability.* As one becomes alert to contingency, one is limited in the claims made and inferences drawn from a given assessment. This expanding sense of humility thus serves as an obligation to those who design assessments and interpret their results to qualify findings even as knowledge about writing studies is expanded. With Gesa Kirsch and Jacqueline Jones Royster, we are committed to the pursuit of human capability as imagined by Nussbaum. Focus on virtues related to humility—compassion, empathy, courage, and commitment—is integral to the new materialist view these virtues adopt in which exceptionalism is abandoned and replaced with affiliation. Seen in this way, figures 8.1 and 8.2 support what Kirsch and Royster reference in their chapter as the deliberate and conscious cultivation of what it means to be human.

- *To continue to advance knowledge, a multidisciplinary perspective is needed to build language-based conceptual systems.* As we hope we have demonstrated in this chapter, our collaboration has resulted in a conceptual model that is beyond that of any single disciplinary perspective. No psychometric scholar will understand the writing construct from the perspective of a teacher of writing, and no writing studies researcher will understand the intricacies of psychometric perspectives. Together, however, one field forces the other to make clear the assumptions and premises behind their models and thus, to the

benefit of students, new systems are developed. These perspectives will require what Patrick W. Berry terms an ethics of listening in his chapter—interrogating deeply held beliefs and listening to the perspectives of others. As John Duffy (2014, 2019) proposed, needed are ethical dispositions in which confirmed beliefs are revisited and expression is considered generously; perhaps an ethics of listening and ethical dispositions are the two most important preconditions for multidisciplinarity, foundations without which harmony between figure 8.1 and figure 8.2 cannot exist. Significantly, we also wish to make explicit that multidisciplinary perspectives allow us to move beyond the lone philosophical voice that establishes a singular ethics agenda. Used wisely, collaborative thinking advances justice and capability. Multidisciplinarity is not only good practice; it is also philosophically justified and morally virtuous.

While much more work remains to be done, these four principles support calls for greater construct validity in our research practices and desires for fairness as an integrative framework for validity and reliability. Perhaps of most significance, these principles support a view of writing assessment as a formative, generative process that can be understood in terms of contextualized, purposive arguments designed, at the end of the day, to advance opportunity.

REFERENCES

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Berkenkotter, Carol, and Thomas N. Huckin. 1993. "Rethinking Genre from a Sociocognitive Perspective." *Written Communication* 10 (4): 475–509.

Chalhoub-Deville, Micheline. 2003. "Second Language Interaction: Current Perspectives and Future Trends." *Language Testing* 20 (4): 369–83. http://dx.doi.org/ 10.1191/0265 532203lt264oa.

Council of Writing Program Administrators, National Council of Teachers of English, and National Writing Project. 2011. *Framework for Success in Postsecondary Writing*. http:// wpacouncil.org/aws/CWPA/asset_manager/get_file/350201?ver=1624.

Cronbach, Lee J., Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam. 1972. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York: Wiley.

De Boeck, Paul, and Mark Wilson. 2004. *Explanatory Item Response Models*. New York: Springer.

Dorans, Neil J. 2011. "Holland's Advice for the Fourth Generation of Test Theory: Blood Tests Can Be Contests." In *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland*, ed. Neil J. Dorans and Sandip Sinharay, 259–72. New York: Springer.

Duffy, John. 2014. "Ethical Dispositions: A Discourse for Rhetoric and Composition." *JAC: A Journal of Composition Theory* 34 (1): 209–37.

Duffy, John. 2019. *Provocations of Virtue: Rhetoric, Ethics, and the Teaching of Writing*. Logan: Utah State University Press.

Elliot, Norbert. 2016. "A Theory of Ethics for Writing Assessment." *Journal of Writing Assessment* 9 (1). http://journalofwritingassessment.org/article.php?article=98.

Gee, James Paul. 1992. *The Social Mind: Language, Ideology, and Social Practice*. New York: Bergin and Garvey.

Gee, James Paul. 2008. "A Sociocultural Perspective on Opportunity to Learn." In *Assessment, Equity, and Opportunity to Learn*, ed. Pamela A. Moss, Diana C. Pullin, James P. Gee, Edward H. Haertel, and Lauren J. Young, 76–108. Cambridge: Cambridge University Press.

Graham, Steve, Alisha Bollinger, Carol Booth Olson, Catherine D'Aoust, Charles MacArthur, Deborah McCutchen, and Natalie Olinghouse. 2012. *Teaching Elementary School Students to Be Effective Writers: A Practice Guide* (NCEE 2012-4058). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education. https://ies.ed.gov/ncee/wwc/Docs/practiceguide/writing_pg_062612.pdf.

Graham, Steve, Julie Bruch, Jill Fitzgerald, Linda D. Friedrich, Joshua Furgeson, Katie Greene, James S. Kim, Julia Lyskawa, Carol Booth Olson, and Claire Smith Wulsin. 2016. *Teaching Secondary Students to Write Effectively* (NCEE 2017-4002). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education. https://ies.ed.gov/ncee/wwc/Docs/PracticeGuide/wwc_secondary_writing_110116.pdf.

Hayes, John R. 2012. "Modeling and Remodeling Writing." *Written Communication* 29 (3): 369–88. http://dx.doi.org/10.1177/0741088312451260.

Holland, Paul W. 2008. "The First Four Generations of Test Theory." Paper presented at the Association of Test Publishers on Innovations in Testing. Dallas, TX, March.

Kant, Immanuel. 2012 [1785]. *Groundwork of the Metaphysics of Morals*. Rev. ed. Trans. and ed. Mary Gregor and Jens Timmerman. Cambridge, UK: Cambridge University Press.

Kelly-Riley, Diane, and Carl Whithaus. 2016. A Theory of Ethics for Writing Assessment. Special issue of *Journal of Writing Assessment* 9 (1). http://journalofwritingassessment.org/archives.php?issue=19.

Leijten, Mariëlle, Luuk Van Waes, Karen Schriver, and John R. Hayes. 2014. "Writing in the Workplace: Constructing Documents using Multiple Digital Sources." *Journal of Writing Research* 5 (3): 285–337. http://dx.doi.org/10.1111/j.1745-3984.1976.tb00178.x.

Lord, Frederic M. 1952. "A Theory of Test Scores." *Psychometric Monograph no. 7*. Princeton, NJ: Educational Testing Service. https://www.psychometricsociety.org/sites/default/files/pdf/MN07.pdf.

Mill, John Stuart. 1879 [1861]. *Utilitarianism*. 7th ed. London: Longmans, Green.

Miller, Carolyn R. 1984. "Genre as Social Action." *Quarterly Journal of Speech* 70 (2): 151–67.

Mislevy, Robert J. 2018. *Sociocognitive Foundations of Educational Measurement*. New York: Routledge.

National Research Council. 2012. *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*. Washington, DC: National Academies Press.

Nussbaum, Martha C. 2011. *Creating Capabilities: The Human Development Approach*. Cambridge, MA: Harvard University Press.

Oliveri, Maria Elena, Robert M. Mislevy, and Norbert Elliot. 2020. "After Admissions: What Comes Next in Higher Education?" In *Higher Education Admissions Practices: An International Perspective*, ed. Maria Elena Oliveri and Cathy Wendler, 347–75. Cambridge, UK: Cambridge University Press. https://doi:10.1017/9781108559607.019.

Poe, Mya, Asao B. Inoue, and Norbert Elliot, eds. 2018. *Writing Assessment, Social Justice, and the Advancement of Opportunity*. Fort Collins and Boulder: WAC Clearinghouse and University Press of Colorado.

Rachaels, James. 1986. *The Elements of Moral Philosophy*. Philadelphia: Temple University Press.

Rawls, John. 1999 [1971]. *A Theory of Justice*. 2nd ed. Cambridge, MA: Harvard University Press.

Schendel, Ellen, and Peggy O'Neill. 1999. "Exploring the Theories of Consequence of Self-Assessment through Ethical Inquiry." *Assessing Writing* 6 (2): 199–227. https://doi.org/10.1016/S1075-2935(00)00008-8.

Slomp, David. 2016. "An Integrated Design and Appraisal Framework for Ethical Writing Assessment." *Journal of Writing Assessment* 9 (2): 1–14. http://journalofwritingassessment.org/article.php?article=91.

Sparks, Jesse R., Yi Song, Wyman Brantley, and Ou Lydia Lou. 2014. *Assessing Written Communication in Higher Education: Review and Recommendations for Next-Generation Assessment*. ETS Research Report no. RR-14–37. Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12035.

Sperber, Daniel. 1996. *Explaining Culture: A Naturalistic Approach*. Oxford, UK: Blackwell.

Suárez, Mauricio. 2004. "An Inferential Conception of Scientific Representation." *Philosophy of Science* 71 (5): 767–79. http://dx.doi.org/10.1086/421415.

White, Edward M., Norbert Elliot, and Irvin Peckham. 2015. *Very Like a Whale: The Assessment of Writing Programs*. Logan: Utah State University Press.

Winner, Langdon. 1980. "Do Artifacts Have Politics?" *Daedalus* 109 (1): 121–36.

Yancey, Kathleen Blake, Linda Robertson, and Kara Taczak. 2014. *Writing across Contexts: Transfer, Composition, and Sites of Writing*. Logan: Utah State University Press.

Zumbo, Bruno D. 2007. "Three Generations of Differential Item Functioning (DIF) Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going." *Language Assessment Quarterly* 2: 223–33. http://dx.doi.org/10.1080/15434300701375832.