Gating Walls and Bridging Gaps: Validity in Language Teaching, Learning, and Assessment

Linda C. Badon[1], Stephen D. Oller[2], Ruixia Yan[3], and John W. Oller, Jr.[4]
University of Louisiana at Lafayette

ABSTRACT

Theoreticians and practitioners often speak as though classrooms and clinics were located outside the real world, but this is not so. The demand for the teaching, learning, and assessment of English language proficiency for international pilots and air traffic controllers is just one practical example touching almost everyone in the world that shows that contexts of actual language use are invariably in the real world just as learners and teachers are. From actual instances of discourse processing, in some cases with life or death consequences, it follows that judgments of validity, like those of ordinary truth, involve the dynamic interactions of persons, sign systems, and variable content in real contexts. Studies of so-called task-based performance assessment (in various forms and by other names) afford many ways to connect teaching, learning, and assessment. The authenticity, representativeness, and consequent generalizability of teaching, learning, and assessment tasks depends on their incorporation of the sign systems, social actions, and realia found in actual contexts of discourse. While codes, contexts, and interactions must be distinguished in theory, in practice they interact holistically. Our theories need to accommodate and account for the synergistic interactions.

INTRODUCTION

On March 27, 1977, the worst air traffic fatality in the history of aviation killed 583 persons. It occurred at Tenerife Airport in the Canary Islands when two passenger-laden Boeings 747 collided on a runway. A Dutch KLM flight was taking off while a Pan Am 747 was crossing the runway. We use this example to argue that some of the walls between language teaching, language learning, and assessment (testing) need to be torn down or else we need to put gates in

[1] L. C. Badon, CCC-SLP, Ph.D., is a licensed speech-language pathologist, child-language specialist, and researcher. Her dissertation and related works suggest that language and literacy instruction directing attention to content are generally superior to methods focusing on sound-letter relations. Correspondence should be sent to Linda C. Badon, UL Lafayette, PO Box 43170, Lafayette, LA 70504-3170. E-mail: badon@louisiana.edu.

[2] Stephen D. Oller, ABD, is a Graduate Assistant and Ph.D. student in the Applied Language and Speech Sciences program at UL Lafayette. He specializes in applications of theoretical semiotics to discourse processing in learning and teaching. His dissertation is about intelligibility of native and nonnative speech. E-mail soller@louisiana.edu.

[3] Ruixia Yan is a Graduate Assistant and Ph.D. student in the Applied Language and Speech Sciences at UL Lafayette. She specializes in measuring proficiency across linguistic and cultural boundaries as well as measurement of language proficiency in international aviation. E-mail rxy3093@louisiana.edu.

[4] John W. Oller, Jr., Ph.D., is Hawthorne Regents Professor at UL Lafayette. He is a specialist in semiotic systems and organizer of the Sertoma International Conference on Autism, spring 2007 in Lafayette, Louisiana. E-mail joller@louisiana.edu.

them. In cases where gaps exist between the activities of teaching, learning and assessment, we believe that some bridges are needed. Our arguments for doing all this are theoretical, and yet they can have profound consequences as our example of the Tenerife accident shows in several ways:

- For one, the accident was evidently caused by misunderstood communications in English, a second language for the pilots and air traffic controllers involved. It could have been prevented with better, more valid teaching, learning, and assessment. Efforts are being made to that end, as we will see later on in this paper.

- For another, the example illustrates essential aspects of the relation between teaching, learning, and assessment tasks and the problem of whether or not any given task is valid — i.e., authentic, representative, and generalizable beyond the situation in which it is used.

- The example gets our attention because we all depend on intelligible language use in international airports by pilots and air traffic controllers. However, the example also illustrates that achieving agreement on the meanings and uses of linguistic representations is central to successful communication in essentially all teaching, learning, and assessment tasks in classrooms and educational contexts. We will argue that comprehensible language use is crucial to teaching, learning, and assessment tasks throughout the real world.

We argue that the validity of any given teaching, learning, and assessment task — whether it is representative, authentic, and generalizable — is just a more complex version of the problem of determining whether a representation of a given state of affairs is true or not. We provide two logical arguments. Both of them show that the construal (production and interpretation) of surface forms of discourse in order to represent faithfully (truthfully) certain changing states of affairs in the real world is the necessary and sufficient basis for any validity to be found in any teaching, learning, and assessment tasks whatever. An essential implication of both arguments taken together is that teaching, learning, and assessment ought to be a lot more closely integrated than they have been in the past. To develop all these ideas, we refer to the kinds of interactions between air traffic controllers and pilots that are needed to accurately represent and manage the location and movements of aircraft arriving and departing from an international airport. This example is useful in showing the importance of truth in representations and validity in language teaching, learning, and assessment.

The first logical argument is derived from Borsboom, Mellenbergh, and van Heerden (2004). They argue for a simple theory of validity grounded in attributes of existing things that differ in ways that cause corresponding differences to arise in measures. The second argument is more general and more powerful. It shows that the conventional signs in any language (including fictions, errors, and deliberate lies) must have been grounded in the past in attributes, events, and states of affairs in the real world in order for them to achieve any intelligibility at all. In the end, we hold that validity in assessment is not just analogous to the simplest mundane sort of truth, but that validity and ordinary truth are one and the same attribute. Their only differences are superficial. Validity is normally regarded as an attribute of tests or measures, while ordinary truth is regarded as an attribute of reports (usually simple assertions) that are consistent with known facts. In order to develop the arguments to be set in place, it is first necessary to provide a context for the discussion. We have chosen the context of international aviation in order to illustrate the nature and force of our key arguments.

LANGUAGE USE AND COMPREHENSION ARE CRUCIAL

Consider the fatal aviation accident on March 27, 1977. In that situation, as in ordinary contexts of communication in general, the truth/validity question depended on how well representations squared with the facts of experience. Einstein (1941/1956) argued that everything in the nature of a theory or representation ultimately depends on how our abstract conceptualizations fit with the sorts of concrete facts that can be known, in part at least, through sensory impressions.

On that fateful day in 1977, there were several factors that made it difficult for the pilots and air traffic controllers to represent the facts faithfully to themselves and each other. One of the key issues was language proficiency. What is meant by certain words and phrases? On the day of the accident in question, sensory impressions were less helpful on account of a heavy ground fog. The controllers in the tower could not see the runway, nor could the pilots of the KLM and Pan Am planes realize their difficulty until seconds before impact. The Dutch pilot could not see the Pan Am aircraft until 13.5 seconds before the collision and the Pan Am pilot did not see the KLM plane coming until about 11 seconds before impact.

The difficulty arose because representations of the air traffic controllers and pilots were not in agreement with the facts. A critical linguistic problem arose when the air traffic controller issued certain climb out and heading instructions to the KLM captain. The KLM captain repeated the instructions and added, “We are now at takeoff.” By this, the captain meant he was accelerating to takeoff speed. He evidently thought he had been cleared for takeoff. The controller understood the statement, “We are now at takeoff,” to mean that the KLM pilot was still waiting at the end of the runway for clearance. The 563 lives lost that day came to depend on whether the phrase “at takeoff” meant “ready and waiting to do so” or “already engaged in doing so.”

Even after the KLM captain committed to takeoff on the false supposition that the elaborate climb-out and heading instructions implied permission to takeoff, the accident still could have been avoided if the air traffic controller had understood the pilot as meaning that he was already taking off. But the controller did not ask for clarification or provide any additional instruction. If the controller had known the KLM plane was already in motion, the controller could have ordered the KLM pilot to abort the takeoff, or he could have ordered the Pan Am pilot to speed up his runway crossing. Or, if the KLM pilot had known that the Pan Am plane was still on the runway, by accelerating faster (and using additional fuel on his takeoff run), the KLM captain could easily have shortened the required distance to lift off and would have been able to clear the Pan Am plane.

Not only is the example relevant to all travelers who fly in or out of international airports, but it is relevant to all communications that depend on shared language uses. A study by Boeing Aircraft showed that of all aviation fatalities during the period from 1982-1991, 11 percent could be attributed in part or in whole to language use problems (Ritter, 1996; Tajima, 2004). The great majority of all aviation incidents, approximately 70%, based on a study of 28,000 safety reports involved problems in "information transfer," and most particularly between pilots and air traffic controllers. Day (2005) wrote, "The most vulnerable link in our ... airspace system is information transfer between air traffic controllers and pilots" (p. 1). Because of the fact that correct understanding of pilots and air traffic controllers through a common language is so critical, the International Civil Aviation Organization (ICAO) has established more stringent English language proficiency standards to be complied with by March 5, 2008 (Mathews, 2004).

On account of the need to ensure safe air travel, especially in international aviation, the ICAO has wisely, we believe, taken steps to reduce the walls and gaps that typically separate the contexts of teaching, learning, and assessment. The new standards they have put in place require stricter compliance with standard phraseology (the language of air traffic controllers and pilots), knowledge of specialized usages within the industry, and knowledge of general (plain) English needed in unusual and unforeseeable circumstances. Although just which unusual situations will arise is unpredictable, it is surprisingly common for unexpected things to happen not only in aviation but in all human experience. For this reason, it is commonly agreed that general proficiency is required over and above special knowledge of aviation terminology in English. Two phases of compliance with new language requirements are envisioned by ICAO: first, testing to ensure adequate English proficiency for the purposes of licensing, and second, retesting to ensure continued maintenance of minimal skills. In the industry in question, near-native skill is required on account of the life and death issues at stake (Mitsutomi & O'Brien, 2004). In establishing tougher assessment and licensing requirements, the aviation industry hopes to create healthy washback effects to the teaching and learning of required language skills.

WALLS AND GAPS ARE EVERYWHERE

In high-stakes settings as well as in language classrooms and clinics in general, traditionally a wall of separation, a gap, or at least a theoretical boundary has been supposed to exist between teaching and testing, and an even greater separation has been supposed to exist between classrooms and clinics and the outside world. Lesser separations abound between the various manifestations, skills, abilities, and dynamic elements of ordinary communication. For instance, distinctions are commonly made between teaching, learning, and assessment tasks aimed at highly focused discrete elements of supposedly distinct components of grammatical systems such as phonology, morphology, lexicon, syntax, semantics, and pragmatics. Distinctions are posited between the content referred to (e.g., the Pan Am plane versus the KLM aircraft), functions of discourse forms ("at takeoff" versus "already taking off"), and performances by certain individuals (e.g., the actions taken by pilots and air traffic controllers in the Tenerife accident). Although it makes excellent sense to distinguish the interlocutors, the surface forms of utterances and actions they deploy (e.g., the words, phrases, and movements), and the content entering into their discursive interactions (e.g., the planes, runways, and unfolding events), to what extent can or should sharp boundaries be imposed between the various elements and interactions?

Referring specifically to oral tasks used in performance assessments of various kinds, Chalhoub-Deville (2001) remarked that "language testers and researchers need to expand their test specifications to include the knowledge and skills that underlie the language construct. Such specifications should be informed by theory and research on the language construct and the language-learning process as well as by systematic observations of the particulars of a given context" (p. 225). We understand this statement to be an argument for a deeper, wider, and richer conception of validity — one that connects testing with learning, and with specific language based performances in particular contexts. We agree with Chalhoub-Deville, and we argue here that the walls of separation that have sometimes been assumed to exist between interpretations, language used, and setting, though theoretically sound in the abstract, have turned out in practice to obscure the essential dynamics of interactions.

However, the traditional separation between teaching and testing is easy to see. It is manifested in separate classes for teaching methods as contrasted with learning. Both of these are usually separated from courses aimed at testing. There are distinct journals for publishing research in these areas, different conventions and organizations, and there is a general though not universal absence of interaction across the areas. Eisner (1999) asserts that educators want more. "We want test scores to tell us about how students address tasks beyond the classroom", and more particularly, we want "valid judgments about 'what they know and can do' in situations that matter" (p. 2). In communication disorders, Westby, Stevens-Dominguez, and Oetter (1996) note that mandated assessments in schools often "provide little useful information to guide intervention" (p. 144). They note that assessments often focus on splinter skills, surface forms of discourse, or bits and pieces of knowledge that have little or no resemblance to the dynamic richness of ordinary contexts of communication.

As a result of such mismatches, a hypothetical wall has been interposed between the language classroom (or the clinic) and the so-called "real" world. For instance, Kim (2004) writes about the need to "generalize about students' ability beyond the learning/testing situation to real-life communication" presupposing the common distinction between the "learning/testing situation" and "real-life communication" (p. 1). Bachman (2002) presupposes the commonly hypothesized separation in saying that a "fundamental aim of most language performance assessments is to present test-takers with tasks that correspond to tasks in 'real-world' settings, and that will engage test-takers in language use or the creation of discourse" (p. 471). Bachman notes that the generalization from teaching, learning, and assessment tasks to the "real" world is commonly regarded as problematic: "The suggestion that assessment tasks may be fundamentally different from pedagogic or, by extension, 'real-life' tasks, not only raises questions about the validity of assessing certain aspects of language ability with certain types of tasks.... but also raises a much more general and perplexing question about the generalizability of research with SLA [second language acquisition] and pedagogic tasks to assessment tasks" (Bachman, 2002, p. 464).

How can the barrier between teaching, learning, and assessment tasks and the larger world beyond be surmounted? There is a growing consensus that one of the most promising approaches to assuring authenticity and generalizability of teaching, learning, and assessment tasks is what has come to be known as performance assessment. Eisner (1999) commented that "performance assessment is the most important development in evaluation since the invention of the short-answer test and its extensive use during World War I" (p. 2). Kim (2004) succinctly summarized about five decades of research on "performance assessment" describing it as "any assessment procedure that involves either the observation of behavior in the real world or a simulation of a real-life activity with raters to evaluate the performance" (p. 1). Kim went on to note that performance assessment is different "from traditional paper-and-pencil tests" because it aims to "get an accurate picture of students' communicative abilities" and more importantly "to generalize about students' ability beyond the learning/testing situation to real-life communication" (p. 1).

There are many variations on the central themes of authenticity and validity that motivate dynamic performance testing in some form or other. Task-based assessment in studies of first and second language acquisition, literacy, and the pragmatics of ordinary communication is perhaps at the center of this paradigm shift in assessment. From about the middle 1980s, "assessment centers" have been widely used in the employment industry by municipalities and other corporate entities. Assessment centers are intensive testing situations grounded in performances on tasks (tests) based on intensive, descriptive, and ethnographic job analyses. They have been used effectively for performance evaluations and promotional assessments (Woehr & Arthur, 2003). Their central objective is to devise tasks that optimally resemble the work on the job in order to ensure validity, but a common finding is that verbal and social skills tend to be the overriding constructs that account for obtained variance (Carless & Allwood, 1997).

In language testing where the focus is intended to be on the language-based proficiencies of individuals as required for various applications, a similar shift toward more and more "real-life"-like tasks has also occurred. In the assessment of communication disorders, despite the recalcitrant use of long-standing so-called "language" tests that are really focused on superficial discrete elements of the phonological, morphological, and lexical forms of speech and writing (Westby et al., 1996), there is a growing interest in dynamic, pragmatic, integrative, descriptive, and more authentic procedures for assessing communication abilities (Damico & Oller, 1991; Goodwin, 1995; Kratcoski, 1998; McNamara, 1996; Westby et al., 1996).

Researchers have noted that widely used standardized tests fail to represent learning and ability to achieve adequately (Y-F. Chen & Martin, 2000). In our increasingly technological world, learners are commonly required to use more complex and subtle forms of thinking than were common prior to the information age. Eisner (1999) argues that learners need to know how to frame problems for themselves, formulate strategies for seeking information, assessing multiple outcomes, considering impact on social relationships, dealing with ambiguity, and changing purposes in the light of new information as it is acquired.

Many traditional assessment tools do not reflect advances in knowledge of language development and learning (Crais & Lorch, 1994; Wetherby & Prizant, 1993). Pierce and O'Malley (1992) reported that traditional forms of assessment do not represent classroom activities. They do not reflect current theories of learning and cognition, nor the abilities students actually need for success. They often focus on superficial elements of discourse products rather than the processes used in creating discourse and inferring interpretations of it. In addition, traditional forms of assessment, especially so-called "standardized" tests, are not well-suited for monitoring progress or informing the school curriculum (Pierce & O'Malley, 1992). We may conclude that traditional approaches to assessment are less authentic, representative, and generalizable as guides for planning appropriate curricula than the teaching, learning, and assessment tasks that are needed (Y-F. Chen & Martin, 2000). We agree with Haynes and Pindzola (1998) that articulating appropriate teaching, learning, and assessment programs requires that the artificial separation of assessment and clinical management be bridged. Assessment should guide clinical intervention and should likewise be guided by accurate appraisals of the knowledge and skills required for success in communication beyond the classroom or clinic (Culatta & Wiig, 2002).

Performance assessment is a movement toward greater authenticity as called for by Rhodes and Shanklin (1993; also Shanklin & Rhodes, 1989). Task-based performance assessment can require learners to demonstrate their knowledge and skills in response to authentic activities (Cooper, 1997). According to Pierce and O'Malley (1992) performance assessment necessarily begins with the observation of actual tasks. It necessarily includes the basis for diagnostic feedback (Hanna, 1993), and the integration of content, dynamic decision-making, and social cooperation (McTighe, Seif, & Wiggins, 2004). In all of these ways, performance assessment is a natural laboratory in which to explore the implications of the reasonable call for more intensive integration in language assessment as called for by Chalhoub-Deville (1996). Using a task-based approach to performance assessment encourages every clinician/teacher to set authentic intervention objectives in the context of authentic tasks. In addition, it requires the integration of task relevant content, skills, knowledge, and the dynamic integration of all these in meaningful discourse.

INTERACTIONS TAKE PLACE IN THE REAL WORLD

Without denying or even questioning the differences at issue between a classroom/clinic and, say, an airport control tower, it remains a fact that such different entities exist in the same real world. Moreover, it is interesting to note that in the radio and telecommunications essential to current international aviation, radiotelephonic conversations may be a more appropriate means of testing than face-to-face interactions. Fischer (2004) has shown that substantial reliability and validity can be attained with scalar evaluations of radiotelephonic interactions when compared against face-to-face interactions and other distance modalities. Similarly, it should be noted that the teaching, learning, and assessment tasks that are used to build or assure the acquisition of language forms, social action skills, or world knowledge are not logically or necessarily different in kind, patterning, or substance from the same elements deployed in contexts beyond the classroom (or any clinical or training setting). In fact, communication settings outside the classroom/clinic may involve the same content, the same patterns of interaction, and some of the same dynamics as the teaching, learning, and assessment tasks in the classroom/clinic.

What is more, in carefully designed research, it is sometimes possible to control, if not to completely eliminate, some of the interacting variables. While this is more difficult when dealing with the ordinary complexities of common communication tasks, it is not entirely impossible, and sometimes factors can be manipulated in such a manner as to orthogonalize their contributions to the difficulty of a particular teaching, learning, and assessment task. What is more, in many ordinary communication contexts, the various sources of complexity are so well controlled and within reach of the performers involved that task performance is, for all intents and purposes, virtually perfect (cf. Xiao & Oller, 1994). In fact, that is what we hope for every time we board an international flight and every time we enter or leave a crowded airspace.

Even the most contrived and artificial teaching, learning, and assessment tasks are situated in the "real" world. We must suppose that many of the same physical, social, and psychological constraints apply in the classroom as in the airport control tower, though perhaps not so unforgivingly. It is better, therefore, to work out the needed skills in classrooms rather than on the runway. In language classrooms, well-designed teaching, learning, and assessment tasks involving high quality sound motion pictures (L. Chen & Oller, in press) can include realistic re-enactments of actual scenarios, conversations, etc., to deploy the same discourse elements with the same phonological, syntactic, morphological, lexical, semantic, and pragmatic constraints imposed upon them as are found in the so-called "real" world. The explicit and implicit instructions necessary to performing teaching, learning, and assessment tasks, moreover, can be made richer, more authentic, and more dynamic than they could ever have been in the old paper and pencil environment of traditional psychological tests (J. W. Oller, Kim, & Choe, 2001). Also, as noted by Greenwood and Rieth (1994) current multimedia technologies enable a much closer integration of teaching, learning, and assessment tasks than has ever been possible before. Indeed, to the extent that the underlying physical, social, and psychological constraints on discourse tasks can be understood and effectively described (J. W. Oller, Chen, S. D. Oller, & Pan, in press), there is no longer any logical basis for supposing they cannot be imported into classrooms.

Skehan (1998) argued for distinctions between: (1) the complexity of the language (code, or the structured forms of discourse) needed to accomplish any particular task, (2) the cognitive complexity of whatever thinking may be required to perform the task, and (3) the level of communicative stress that may be involved in performing the task. However, Bachman (2002) saw two problems in such an analysis. For one, the abilities of different performers are confounded with teaching, learning, and assessment tasks, and for another, Bachman sees it as a mistake to treat task difficulty as an independent variable. He argues that code complexity is the only factor independently associated with teaching, learning, and assessment tasks and that the other variables involve interactions between distinct elements. For instance, cognitive complexity and communicative stress are believed to depend on interactions of the performer with the task. Cognitive complexity is supposed to be a function of processing and familiarity which are variable attributes of performers. Agreeing with Norris, Brown, Hudson, and Yoshioka (1998; also see Norris, Brown, Hudson, & Bonk, 2002), Bachman sees communicative stress as partly dependent on how well performers can handle the task(s). This depends on the language proficiency, willingness to take risks, and how different individuals work through any given task. Bachman further argues that attempts to measure or predict the difficulty of distinct teaching, learning, and assessment tasks has been notably unsuccessful. He concludes that the relevant research shows "virtually no systematic relationship among a priori estimates of difficulty based on difficulty factors and empirical indicators of difficulty" (p. 463). For the foregoing reasons, he concludes that we should view "tasks as sets of characteristics, rather than as holistic entities" and that we should distinguish "task characteristics ... that require no assumptions regarding test-takers or how they may interact with the task" from "attributes of test-takers" and the latter from "interactions between test-takers and task characteristics." Finally, Bachman argues that we should think of "interactions as interactions" (p. 469).

A telling example is the difficulty associated with a physical task like high jumping. Bachman (2002) points out that for one jumper the difference between a bar height of 5'8" versus 5'10" would be a lot, but for another jumper, someone able to clear 6'4" or more, the 2" difference between 5'8" and 5'10" would be relatively insignificant. On this basis, Bachman supposes that "difficulty does not reside in the task alone, but is relative to any given test-taker" (p. 462). More importantly, as Bachman argues, the interaction between relative abilities of learners and teaching, learning, and assessment tasks "has clear implications for the way in which we interpret and use test results, and for the validity of these interpretations and uses" (p. 468). We agree that the interaction between teaching, learning, and assessment tasks and the abilities of performers (test-taker/learners) is crucial to the validity of such tasks, but we present two arguments to show conclusively, we believe, that the tasks to be performed are logically prior to and the basis for judgments about the relevant knowledge, skills, and abilities of performers.

In presenting these arguments, we offer what we believe is a more complete, integrated, and consistent definition of validity. The first part of our argument is drawn from Borsboom et al. (2004). They argue for a simple theory of validity grounded in existing things and attributes that differ in ways that cause corresponding differences to arise in measures. The second part of our argument is more general. It depends on nothing but the nature of the conventional signs of any language. Such signs, as Peirce (1903/1934) was the first to clearly show, invariably consist of three aspects in their most complete manifestations: a concrete aspect grounded in material facts (shown mainly in icons), an indexical aspect involving movements that connect observers with the material world (that link distinct icons), and an abstract symbolic aspect that is essentially conceptual (the abstract linguistic aspect).

GROUNDING TASKS AND MEASURES

Borsboom et al. (2004) propose arguments grounded in the propositions that "a test is valid for measuring an attribute if and only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure" (p. 1061). They illustrate their premise by saying that "we cannot see how the sentences *Test X measures the attitude toward nuclear energy* and *Attitudes do not exist* can both be true" (p. 1064). On the other hand, if something to be measured actually exists, they argue a valid measure requires a direction of causation from the thing to be measured to variations in the measure. As a result, validity is directional. "The direction goes from the world to psychologists' instruments" (p.1066). The validity question is simply "whether the attribute to be measured produces variations in the measurement outcomes" (p. 1069). This view differs from approaches that attempt to ground validity in the correlations between similar or diverse measures (e.g., Cronbach & Meehl, 1955). Borsboom et al. argue that any "conceptualization of validity in terms of covariation, rather than causality, is flawed" (p. 1066).

A critical demonstration, that the covariation idea comes up short, flows from theoretically perfect measurements of length. For instance, consider perfect measures applying to objects of equal length. The correlations between such measures will be zero for want of variability and yet the measures may be exactly accurate and valid. Borsboom et al. (2004) are not arguing that correlations are not useful, by any means, but that logically they cannot fulfill the requirements of validity. Unless there is some other basis for hypotheses about "what happens between item administration and item response, then one will find no clarity in tables of correlation coefficients" (p. 1063), and even the powerful tool of confirmatory factor analysis (Jöreskog, 1971) must come up short of the mark. The upshot is that validity cannot be assured by mere examinations of correlation coefficients or any higher procedures grounded in correlations alone. This is not to say that correlation coefficients should not be examined, nor that Cronbach and Meehl (1955) multi-trait multi-method matrices should not be examined, but that at some point, some of the measures entered into the analyses must be associated with actual processes performed by existing persons in the contexts of teaching, learning, and assessment tasks that have independent reality as well as independent claims to authenticity, representativeness, and generalizability. Otherwise, convergences and divergences between measures are insufficient by themselves to enable secure claims of validity.

Next, we show an independent basis for the conclusion reached by Borsboom et al. (2004), and we show that their essential conclusion logically accommodates rather than refutes

Messick's arguments that the interpretations and social consequences of tests (Messick, 1989, 1994, 1998), or in our case teaching, learning, and assessment tasks, are also relevant to the judgment of their validity. By producing our more general logical basis for grounding validity in the common sense notion of ordinary truth, we also show that Messick's arguments remain compatible rather than in contrast with the claims of Borsboom et al. For instance, if teaching, learning, and assessment tasks can be devised that demonstrably enable acquisition of the skills necessary to reduce aviation accidents and fatalities (as the ICAO aims to do; see Day, 2004, p. 22), we would take this as prima facie evidence that those tasks were accomplishing their intended functions, i.e., that they were valid. Day says, "Improving communication effectiveness is one of the few areas where a significant positive safety impact is possible at an affordable cost and effort" (2004, p. 22).

A MORE GENERAL LOGICAL ARGUMENT

Our argument showing that authentic, representative, and generalizable teaching, learning, and assessment tasks are theoretically attainable is profoundly simple. However, owing to the fact that it is based on nothing but the conventionality of the fully abstract and general symbols of a natural language, it is abstract. It derives from studies of theoretical semiotics where it is clear that the essential virtue of conventional signs — that is of the words, phrases, and clauses of discourse in any natural language and in any real context of social action constrained by rational purposes — is their nearly absolute and completely necessary generalizability (Peirce, 1903/1934; J. W. Oller, 2005; S. D. Oller, in press).

Take an obvious generalization that applies to and that can be derived from the Tenerife accident. Two massive objects moving into the same space will collide with destructive consequences. What is less obvious to the casual observer is that the generalization just stated depends for its sense on the association of abstract representations validly (truly and according to their normal conventional applications) with real things in the material world through one or more natural languages. While thousands of pictures of the Tenerife accident and its aftermath exist, to identify any of those pictures initially as distinct from pictures of other accidents, words in some form or other are required. To reasonably infer what happened on that day, we require access to the discourse in which the key interlocutors were engaged. The language of interaction between pilots and controllers was English although the airport was located in a Spanish speaking community and neither the Dutch pilot nor either of the air traffic controllers directing the planes that collided spoke English as their native language. With the assistance of icons, indexes, and symbols showing us the movements of the aircraft we can visualize the scene as it may have appeared to the KLM pilot when he first saw the Pan Am plane in his path as it emerged from the fog. Likewise, we can visualize the approaching KLM aircraft from the viewpoint of the Pan Am pilot as the looming hulk bore down on the barely moving Pan Am plane. By analyzing the verbal exchanges as recorded during the last moments of both flights, much of what went wrong can be reconstructed.

Consider, however, that for us now interacting as writers with each other and with our readers as consumers of this text, in order for any of us to construct representations of the planes and the 583 persons they contained prior to impact, we are almost absolutely and exclusively dependent on the conventional signs of a shared language. Since we were not there, we only know about the accident through the conventional symbols of a common language. We only know of the accident through linguistic representations. Films or animations, even realistic motion pictures with sound showing the collision and its aftermath, would be impossible to associate for certain with the right time, place, and persons, except for identifying conventional signs, referring phrases, and linguistic forms appropriately associated with the material objects and their actual space-time contexts. For precisely the same reasons, within extremely wide and flexible limits, it is the nature of the general linguistic signs of a natural language to enable access not only to the context of the accident on March 27, 1977 at Tenerife airport, but to any context whatever in any setting.

In fact, abstract conventional symbols (words, phrases, and sequences of them) are constrained only by the effort required to connect those symbols by inference through indexes (utterances, gestures, and social acts) to icons (e.g., bodily persons, objects, pictures, printed words, and the like) showing the referents of those symbols. For instance, to understand fully what happened at Tenerife airport, the phrase "Tenerife airport" needs to be associated with an island in the Canaries, south of Spain. The symbols referring to the KLM flight need to be connected with a particular aircraft on that day. In other words, the essential problem of true representation in ordinary discourse involves the conventional association of abstract symbols (the words of some language) through indexes (articulate gestures, especially the utterances of speech or the articulate movements of writing) with actual persons, events, and contexts (things and states of affairs that we or others have known through their senses). The various material contexts of experience are the sort that might be represented in high quality sound motion pictures (iconically).

All the foregoing shows why it is essential for valid teaching, learning, and assessment tasks to make use of all the essential elements in the dynamic contexts of interaction. The boundaries between interlocutors, content, and forms of language though real enough are often actually crossed. The unity of coherence that is often shared by distinct interlocutors is sufficient to show that real separations between persons, things, events, and contexts can often be surmounted if not entirely removed. Airplanes commonly take off from one location and land in another without any accidents. In order for this to happen pilots and air traffic controllers need to agree with each other and with the relevant facts. For this to occur, the boundaries between persons, contexts, and material facts must be bridged. The walls and gaps are effectively crossed in successful communications. The fact that linguistic barriers posed by distinct languages can also be crossed is shown in high quality translation across languages (Peirce, 1903/1934; Xiao & Oller, 1994). Similarly, the genuine generalizability of discourse forms across distinct contexts is also shown in every successful instance of first or second language acquisition. The fact that such things do occur and commonly result in relatively complete success is seen in the fact that planes usually reach their destinations and collisions in crowded international airspaces and on the ground are usually avoided.

While classrooms and airports are usually different in important ways, relations between words and their conventional associations with referents (persons, events, settings, and dynamic interactions over time) are sufficiently general that the discursive processes that take place in a classroom need not be different at all in crucial respects from ones that take place elsewhere. In fact, there is no reason to suppose that the classroom cannot be transported to the airport, or to a moving airplane, a control tower, or wherever. More importantly, the discourse that takes place in any setting whatever can necessarily be generalized to all similar settings exactly to the extent of the similarity between the settings relative to the conventional signs deployed. At the theoretical limit of identity (where things are exactly the same) the generalization is perfect (complete). At the opposite theoretical limit of complete dissimilarity (where things have no similarities at all) no generalization whatever is possible. For instance, the factors producing the collision at Tenerife Island can be generalized to all other material contexts exactly to the extent of the similarity between the Tenerife context and the others to which we wish to generalize. It is for this reason, and this reason only, that agencies involved in the aviation industry are committed to examine every aviation fatality, and every near miss, on the theory that factors contributing to such incidents are generalizable and that by studying them closely similar incidents can be prevented from occurring in the future.

In the important industry of international aviation, the ultimate test of whether or not teaching, learning, and assessment have been made more valid is whether or not international aviation has become safer or not. The extent to which the former walls and gaps separating teaching, learning, and assessment have been broken down, gated, or bridged in the aviation industry can presumably be judged in part in terms of whether or not near misses, accidents, and fatalities in international aviation have been reduced. This is a stringent empirical requirement, but according to Day (2005) the first half of 2004 was the safest ever in the history of aviation. Perhaps, the stringent validity requirement is being met better now than before. If this is so, that is, if aviation is made safer by tightening assessment and licensing requirements, it would seem that these same actions have also helped to bring language teaching and learning more closely into line. Also, in using this argument and showing it as a validity claim we knock down some additional walls that would keep validity from extending to consequences in the real world. On the contrary, we believe that improved teaching, learning, and assessment should have measurable effects on air traffic safety (at least to the extent that other factors can be held constant).

CONCLUSION

How do we know that well-designed teaching, learning, and assessment tasks are generalizable? They must be to the extent that their linguistic signs conform to the relevant conventions of discourse. To that extent, they must generalize to all the contexts where those same conventions are respected. This consequence is necessarily entailed by the very nature of every conventional sign. The very existence of conventions which are essential to the existence of meaningful abstract linguistic symbols, and therefore to the existence of any languages whatever, absolutely requires and entails generalizability across distinct contexts of communication. Finally, we conclude that once the artificial boundary between the abstract realm of constructs and the real world of hard objects, e.g., bodily persons and vehicles that sometimes collide, is gated enabling a flow of material content to be abstracted from actual referents, their attributes, and their interactions to representations, theories, and measures, Messick's argument (1989, 1994, 1998) for interpretability and felicitous (ethical) consequences of test use also extends to teaching, learning, and assessment tasks in general. Although the direction of flow from teaching, learning, and assessment tasks must logically be from the task in the real world to abstract constructs and representations, there is no sound reason to suppose that authentic teaching, learning, and assessment tasks cannot be devised so as to be representative and generalizable to countless other tasks that may be encountered elsewhere, even in the "real" world. The reverse, however, does not hold. Although we can change things in the real world to a small degree through bodily movements and effortful action, just saying something doesn't make it so. The direction of flow is from the concrete to the abstract. For that very reason, the mere interpretability of teaching, learning, and assessment tasks, (as Messick argued) along with desired (and demonstrated) social consequences, can be used to infer the validity of those tasks.

REFERENCES

Bachman, L. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*, 453-476.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061-1071.

Carless, S. A., & Allwood, V. E. (1997). Managerial assessment centers: What is being rated? *Australian Psychologist, 32*, 101-105.

Chalhoub-Deville, M. (1996). Test interpretation, test use and pedagogical implications. *Australian Review of Applied Linguistics, 13*, 188-207.

Chalhoub-Deville, M. (2001). Task-based assessments: Characteristics and validity evidence. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 210-228). Essex, UK: Pearson Education.

Chen, L., & Oller, J. W., Jr. (in press.) High quality sound motion pictures in L2 curricula: Why and how they work. *Canadian Modern Language Review*.

Chen, Y-F., & Martin, M. A. (2000). Using performance assessment and portfolio assessment together in the elementary classroom. *Reading Improvement, 37*, 32-40.

Crais, E. R., & Lorch, N. (1994). Oral narratives in school-age-children. *Topics in Language Disorders, 14*, 13-28.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Cooper, J. D. (1997). *Literacy: Helping children construct meaning* (3rd ed.). Boston, MA: Houghton Mifflin.

Culatta, B., & Wiig, E. H. (2002). Language disabilities in school-age children and youth. In G. H. Shames & N. B. Anderson (Eds.), *Human communication disorders: An introduction* (6th ed., pp. 218-257). Boston, MA: Allyn & Bacon.

Damico, J. S., & Oller, J. W., Jr. (1991). Theoretical considerations in the assessment of LEP students. In E. V. Hamayan & J. S. Damico (Eds.), *Limiting bias in the assessment of bilingual students* (pp. 77-110). Austin, TX: Pro-Ed.

Day, B. (2004). Heightened awareness of communication pitfalls can benefit safety. *ICAO Journal, 59*, 20-22.

Day, B. (2005). ICAO standards and recommended practices: An overview. Retrieved January 25, 2005, from http://www.icao.int/icao/en/anb/meetings/IALS/proceedings/PAPERS/2-Day.pdf.

Einstein, A. (1956). The common language of science. In Author, *Out of my later years* (pp. 111-113). Secaucus, NJ: Citadel. (Originally a radio talk in 1941).

Eisner, E. (1999). The uses and limits of performance assessment. *Phi Delta Kappan, 80*, 658-660.

Fischer, D. C., Jr. (2004). *Comparing face-to-face in distance modalities in conducting Arabic and Russian speaking proficiency tests*. Unpublished doctoral dissertation, University of New Mexico, Albuquerque.

Goodwin, C. (1995). Co-constructing meaning in conversations with an aphasic man. *Research on Language and Social Interaction, 28*, 233-260.

Greenwood, C. R., & Rieth, H. J. (1994). Current dimensions of technology-based assessment in special-education. *Exceptional Children, 61*, 105-113.

Hanna, G. S. (1993). *Portfolios and beyond: A guide to implement*. Thousand Oaks, CA: Corwin.

Haynes, W. O., & Pindzola, R. H. (1998). *Diagnosis and evaluation in speech pathology* (5th ed.). Boston, MA: Allyn & Bacon.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109–133.

Kim, H. (2004). Task-based performance assessment for teachers: Key issues to consider. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics, 4(*2). Retrieved January 20, 2005, from http://www.tc.columbia.edu/tesolalwebjournal/forumFall2004.htm.

Kratcoski, A.M. (1998). Guidelines for using portfolios in assessment and evaluation. *Language, Speech and Hearing Services in Schools, 29*, 3-10.

Mathews, E. (2004). New provisions for English language proficiency are expected to improve aviation safety. *ICAO Journal, 59*, 4-6.

McNamara, T. (1996). *Measuring second language performance*. London: Longman.

McTighe, J., Seif, E., & Wiggins, G. (2004). You can teach for meaning. *Educational Leadership 62*, 26-30.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*, 13–23.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research, 45*, 35–44.

Mitsutomi, M., & O'Brien, K. (2004). Fundamental aviation language issues addressed by new proficiency requirements. *ICAO Journal, 59*, 7-9, 26, 27.

Norris, J. M., Brown, J. D., Hudson, T. D., & Bonk, W. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing, 19*, 395–418.

Norris, J. M., Brown, J.D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments*. (Vol. SLTCC Technical Report 18). Honolulu, HI: Second Language Teaching and Curriculum Center, University of Hawaii at Manoa.

Oller, J. W., Jr. (2005). Common ground between form and content: The pragmatic solution to the bootstrapping problem. *Modern Language Journal*, *89*, 92-114.

Oller, J. W., Jr., Chen, L., Oller, S. D., & Pan, N. (2005). Empirical predictions from a general theory of signs. *Discourse Processes, 40*, 115-144.

Oller, J. W., Jr., Kim, K. & Choe, Y. (2001). Can instructions to nonverbal IQ tests be given in pantomime? Additional applications of a general theory of signs. *Semiotica, 133*, 15-44.

Oller, S. D. (in press). Meaning matters: An application of a general theory of signs to language intervention. *Journal of Communication Disorders*.

Peirce, C. S. (1934). Pragmatism and pragmaticism. In C. Hartshorne & P. Weiss (Eds.), *Collected papers of Charles Sanders Peirce* (Vol. 5, pp. 13-131). Cambridge, MA: Harvard University Press. (Original work published in 1903)

Pierce, L. V., & O'Malley, J. M. (1992). *Performance and portfolio assessment for language minority students. Program information guide series, 9*. (ERIC Document Reproduction Service No. ED346747)

Ritter, J. (1996, January 9). Transcript of crash shows controller error/Review reveals poor English "Over and Over." *USA Today,* p. 7A.

Rhodes, L. K., & Shanklin, N. L. (1993). *Windows into literacy: Assessing learners K-8*. Portsmouth, NH: Heinemann.

Shanklin, N. L., & Rhodes, L. K. (1989). Transforming literacy instruction. *Educational Leadership, 46*, 59-63.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Tajima, A. (2004). Fatal miscommunications: English in aviation safety. *World Englishes, 23*, 451-470.

Westby, C.E., Stevens-Dominguez, M., & Oetter, P. (1996). A performance/competence model of observational assessment. *Language, Speech, and Hearing Services in Schools, 27*, 144-156.

Wetherby, A., & Prizant, B. (1993). Profiling communication and symbolic abilities in young children. *Journal of Childhood Communication Disorders, 15*, 23-32.

Woehr, D. J., & Arthur, W. (2003). The construct-related validity of assessment center ratings: A review and meta-analysis of the role of methodological factors. *Journal of Management, 29*, 231-258.

Xiao, S-Y., & Oller, J. W., Jr. (1994). Can relatively perfect translation between English and Chinese be achieved. *Language Testing, 11*, 267-289.