# If Language is a Complex Adaptive System, What is Language Assessment?[1]

Robert J. Mislevy
University of Maryland

Chengbin Yin
Center for Applied Linguistics

## Abstract

Individuals' use of language in contexts emerges from second-to-second processes of activating and integrating traces of past experiences—an interactionist view compatible with the study of language as a complex adaptive system (LaCAS), but quite different from the trait-based framework through which measurement specialists investigate validity, establish reliability, and ensure fairness of assessments. This article discusses assessment arguments from an interactionist perspective. We argue that the familiar concepts and methods of assessment that evolved under a trait perspective can be gainfully reconceived in terms of the finer-grained perspective of interactionism, and illustrate how key ideas relate to familiar practices in language testing.

Key words: Assessment arguments, LaCAS, language testing, interactionist perspective.

---

[1] Presented at "Language as a Complex Adaptive System," an invited conference celebrating the 60th anniversary of *Language Learning*, at the University of Michigan, Ann Arbor, MI, November 7-9, 2008. Will appear in the special anniversary issue of *Language Learning* in 2009. The first author's work was supported by a grant from the Spencer Foundation.

# Introduction

An educational assessment embodies an argument from what we see people say, do, or make in a handful of particular situations, to inferences about their capabilities as more broadly construed. Although the visible elements of assessments such as tasks, scoring rubrics, and measurement models are familiar, it is a conception of capabilities that shapes their form and gives them meaning. Different conceptions give rise to markedly differently requirements for what to observe and how to interpret it, in order to support markedly different claims about examinees (Mislevy, 2003, 2006; Mislevy, Steinberg, & Almond, 2003). This article considers implications for assessment of a conception of individuals' language capabilities from an interactionist perspective, a view aligned with research on language as a complex adaptive system (LaCAS; Beckner, et al., in press).

The following section reviews key ideas of interactionism, focusing on aspects that become important in assessments meant to support learning, evaluate capabilities, or make decisions about individuals' capabilities with respect to a given language. A framework for assessment arguments is then reviewed. Ways the interactionist perspective impacts assessment arguments and assessment use arguments are then discussed, drawing on current work in language testing (e.g., Chalhoub-Deville, 2003, Chapelle, 1998, Douglas, 1998, 2000).

# LaCAS and an Interactionist Approach to Language Testing

> Language as a CAS [complex adaptive system] involves the following key features: The system consists of multiple agents (the speakers in the speech

> community) interacting with one another. The system is adaptive, that is, speakers' behavior is based on their past interactions, and current and past interactions together feed forward into future behavior. A speaker's behavior is the consequence of competing factors ranging from perceptual constraints to social motivations. The structures of language emerge from interrelated patterns of experience, social interaction, and cognitive mechanisms. (Beckner, et al., in press)

Studies of language as a complex adaptive system are an important contributor to an emerging integration of individual, situative, and social perspectives on cognition (Gee, 1992; Greeno, 1998). Although a review of this work is beyond the scope of the present article, this section summarizes key ideas and offers terminology in order to ground a discussion of language testing. There is broad agreement on core set of ideas:

- Performances are comprised of complex assemblies of component information-processing actions that are adapted to task requirements during performance. These moment-by-moment assemblies build on past experience, and incrementally change capabilities for future action.
- A connectionist paradigm proves useful to frame the intrapersonal processes of learning, memory activation, and situated action. Not coincidentally, this paradigm reflects neuropsychological research on the mechanisms through which brains accomplish these processes.
- The patterns through which people in communities interact shape individuals' learning, in all its linguistic, cultural, and substantive aspects. Intrapersonal

learning is becoming attuned to extrapersonal patterns, in order to perceive, act in, and create situations.

- The extrapersonal patterns themselves evolve over time as individuals use them, extend them, and recombine them in novel ways. The connectionist paradigm can also be gainfully applied to study adaptations at this level.

This research holds profound implications for language testing, for it is from this perspective we would want to design and use assessments. Language testing researchers are now seeking to develop conceptions and methodologies for language assessment that build from the start around the dynamic interaction among people, capabilities, and situations--what is called an *interactionist* approach to language testing (Bachman, 2007; Chalhoub-Deville, 2003; Chapelle, 1998). We draw this research, as well as work in allied fields. Strauss and Quinn's (1997) cognitive theory of cultural meaning, Kintsch's (1998) construction-integration (CI) model of reading comprehension (representative from among many sources), and Young's (2002, 2008) notion of developing resources to participate in discursive practices prove useful for framing points in the discussion.

Strauss and Quinn (1997) highlight the interplay between internal and external realms in the production of cultural, or extrapersonal, meanings. Cultural meanings, i.e., typical interpretations of an event or an object as invoked in a person, are created and maintained in such interactions in a community, the outcomes of which entail durability but allow for variation and change. Meanings are intersubjectively shared, in that people develop mental structures that, while unique in each individual, share salient aspects as their

interpretation of the worlds develops under similar circumstances or as they need to communicate about shared experiences.

Kintsch's (1998) work dovetails with and further extends Strauss and Quinn's with regard to intrapersonal cognition. Kintsch represents knowledge as an associative net and relevant concepts as nodes. The *construction* processes activate an inconsistent, incoherent associative network on the basis of linguistic input, contextual factors, and the comprehender's knowledge base. Mutually reinforcing elements strengthen and isolated elements deactivate in the *integration* processes. The resulting "situation model" mediates understanding and action, akin to Fauconnier and Turner's (2002) notion of a "blended mental space." Kintsch argues that the CI paradigm applies more generally, so that an analog of a situation model would build from cues and patterns (or models, as we will call them here) at all levels: linguistic, in terms of phonology, grammar, and conventions, and pragmatic, cultural, situational, and substantive[2]. We will use the abbreviation "L/C/S models" to stand collectively for linguistic, cultural, and substantive models.

More specifically in the context of language, Young (2000, 2008) describes talk activities that people do as "discursive practices," the construction of which depends on a conglomeration of linguistic, interactional, and identity resources that enable "the ways in which participants construct interpersonal, experiential, and textual meanings in a practice" (Young, 2008, p. 71). The Strauss and Quinn and the Kintsch frameworks help

[2] By substantive, we mean the knowledge and activity patterns related to the substance of an activity, such as cars, cooking, video games, interpersonal relationships, or any of the myriad domains that people engage in.

us understand the nature of such resources. Developing them and being able to bring them to bear in appropriate situations is the goal of learning, and what assessments are to provide information about (that is, the inferential targets of assessment).

## The Structure of Assessment Arguments

Explicating assessment as evidentiary argument brings out its underlying structure, clarifies the roles of the observable elements and processes, and guides the construction of tasks (Kane, 2006; Messick, 1994). The particular forms of the elements and processes will be shaped by a perspective on the nature of the knowledge, skills, or capabilities that are to be assessed. This section sketches that structure, and, anticipating the contrast with assessment from an interactionist perspective, illustrates it with a simple example from discrete-points language testing.

Figure 1 is a depiction Toulmin's (1958) schema for evidentiary reasoning. The *claim* (C) is a proposition we wish to support with *data* (D). The arrow represents inference, which is justified by a *warrant* (W), a generalization that justifies the inference from the particular data to the particular claim. Theory and experience provide *backing* (B) for the warrant. In any particular case we may need to qualify our conclusions because of *alternative explanations* (A) for the data.

[[Figure 1]]

Figure 2 applies these ideas to assessment arguments (Mislevy, 2003, 2006). At the top appears a claim, justified by assessment data through a warrant. At the bottom is a student's action in a situation: The student says, does, or makes something, possibly extending over time, possibly interacting with others. Note that it is interpretations of the

actions rather than the actions themselves which constitute data in an assessment argument, and these interpretations are themselves inferences to be justified by warrants cast in a conception of knowledge and its use.

[[Figure 2]]

Figure 3 elaborates this region of the argument to bring out the interactive and evolving nature of situated action, to accommodate data that arise in more complex assessments. The situated performance is represented as a spiral through time, with arbitrary segments depicted (e.g., conversation turns, discourse topics, speech acts), because as the examinee (inter)acts, the situation evolves. The spiral indicates that the examinee's actions—and, if other people are involved, their actions too, as represented by the gray rectangles—play a role in creating the situation in each next instant.

[[Figure 3]]

The assessment argument encompasses three kinds of data:

- Aspects of the person's actions in the situation,
- Aspects of the situation in which the person is acting, and
- Additional information about the person's history or relationship to the situation.

The last of these does not show up directly in the formal elements and processes of an assessment. We will see the crucial role it plays in assessment from an interactionist perspective.

Figure 4 augments the assessment argument with an assessment use argument (Bachman, 2003). The *claim* of the assessment argument is *data* for the use argument. The claim of the use argument serves some educational purpose such as guiding learning,

evaluating progress, or predicting performance in future (criterion) situations. In language testing, the criterion typically concerns the use of language in particular kinds of real-world situations, or target language usage (TLU; Bachman & Palmer, 1996).

[[Figure 4]]

The simple example from a discrete-points grammar test illustrates these ideas. An examinee is asked to supply the form of an English verb in a sentence:

John _____ the apple yesterday. [eat]

Context is minimal and performance is simple; the task is neither interactive nor purposeful. The task does nevertheless provide sufficient semantic and syntactic information to determine the targeted verb form (i.e., the data concerning the situation), and activate an aspect of English usage that is relevant in language use situations. An examinee's response of "ate" rather than "eated" provides evidence (i.e., the data concerning the performance) for a claim about the examinee's capabilities with regard to this point of grammar, although we shall see that its force is limited under an interactionist perspective.

## Assessment Arguments through an Interactionist Lens

Characteristics of an interactionist perspective that hold implications for assessment are the interactive, constructive nature of language in use; the joint dependence of communication on many patterns, of many kinds, at many levels; and the dependence on previous personal experience to form long-term memory patterns that can be activated in relevant situations. Particular attention is accorded to the features of assessment

situations and criterion situations in terms of L/C/S models they "should" evoke, in terms of the assessment purpose.

**Illustrative Assessment Configurations**

To ground a discussion of assessment from an interactionalist perspective, we will consider the four typical configurations used in language testing listed below. These are not meant to be an exhaustive set of configurations or a taxonomy; they have been selected from among many possible assessment configurations because they are widely used and because they can be used to illustrate key issues in language testing from an interactionist perspective.

(C1) *Lean context, predetermined targets, to support learning.* Discrete-points tasks like "John _____ the apple yesterday" can be used to support instruction under the behaviorally cast "focus on forms" approach.

(C2) *Lean context, predetermined targets, for broad purposes.* Traditional language proficiency tests are large-scale, context-lean, minimally-interactive, highly constrained by costs and logistics, and are meant to support broad and vaguely defined inferences. Examples include TOEFL and the Defense Language Proficiency Tests (DLPTs).

(C3) *Rich context, predetermined targets, for focused purposes.* This configuration is meant to support inferences about examinees' performance in more specified contexts. For example, the International Teaching Assistants Evaluation (ITAE) is used to qualify non-native English speakers as teaching assistants at the University of Maryland.

(C4) *Rich context, opportunistic targets, to support learning.* Task-based language tests can be used in instructional settings to provide feedback and guide instruction for learners based on their unfolding performances. This strategy is related to the "focus on form" instructional approach. ITEA tasks can also be used in classroom to prepare students for the qualification examination.

A brief preliminary discussion of the first and second configurations from the perspective of behavioral and trait approaches to assessment follows immediately. New considerations that an interactionist perspective raises are then addressed, and illustrated in further discussion of the second configuration and of the third and fourth configurations.

**Configuration 1: Lean context, predetermined targets, to support learning**

The previous example with discrete-points language testing illustrates this configuration. Performance in a lean context provides evidence for determining whether the student has acquired the capability to carry out such tasks in previous equally-lean and decontextualized learning situations, or would benefit from additional practice. While the behaviorist pedagogical approach can be debated, the "lean context, predetermined targets, to support learning" configuration clearly illustrates a coherent coupling of an assessment and a purpose.

**Configuration 2: Lean context, predetermined targets, for broad purposes**

In this configuration the assessment context is again lean, but the contexts of use for which inferences are intended are complex though vaguely specified: a variety of real-world language use situations as might be encountered in higher education in a North

American university in the case of TOEFL, or in the case of the DLPTs, a range of settings encountered by military personnel in country, foreign service workers doing diplomatic or bureaucratic functions, or intelligence analysts. Every situation within these criterion settings draws upon not only linguistic but also cultural, social, and substantive patterns. Each is interactive in its own way, each involves its own goals and actors, and there are great variations among them.

For any particular criterion situation, then, much that determines success in that situation, in terms of other L/C/S models the examinee would need to bring to bear, will not be present in the assessment situation. Some make successful use of the targeted language models *more* likely (e.g., affinities with the situations in which they were learned and practiced), while some make successful use *less* likely (e.g., unfamiliar situations, high stress). The limitation arises from applying the same assessment evidence to all use situations—even when from a formal view it is common to them all—when it is stripped of the context of any of them.

From the perspective of educational and psychological measurement, the problem is *construct under-representation* with respect to any particular use situation. From the perspective of assessment arguments, it raises alternative explanations for both good and poor assessment performance as a prediction of performance in the use situations. From the interactionist perspective, the context lean assessment situations do not adequately reflect the language use situations that are of ultimate interest.

## Explicating the assessment argument

"Person acting within situation" is the fundamental unit of analysis from an interactionist perspective (Chalhoub-Deville, 2003). Traditional assessment places the person in the foreground, in terms of theories of traits, although task design and response evaluation have always been integral to the enterprise. Viewing assessment in terms of interacting elements in an argument emphasizes their interconnectedness and co-equal status. Every element in an assessment argument, and every relationship among elements, can be viewed in terms of L/C/S models, to understand and guide the design and use of a test. Even though much of the machinery and methodology that evolved to serve assessment under the trait and behavioral perspectives can still be used, it can be reconceived in interactionist terms. It will not generally be feasible to comprehensively model all aspects of examinee capabilities, contextual features, and performances in full detail in practical assessment applications. This is as it should be, however, since models of perfect fidelity are not necessary for applied work at a higher level of analysis. As in any application of model-based reasoning, the question is not whether the model is a faithful representation of a phenomenon, but whether the simpler modeled representation captures the right aspects for the purpose and under the constraints of the job at hand (Mislevy, in press).

A claim in any assessment argument is based on observing a person act within a few particular situations, construed in a way meant to hold meaning beyond those situations. A claim cast in trait terms such as "control of grammar," "fluency," or "reading comprehension" foregrounds a characteristic of an individual, posited to characterize regularities in behavior over a range of situations (Messick, 1989) and hides the

dependence on context. Reliability and generalizability studies first explore situations more or less like the testing situation and variations in examinees' actions across them (Cronbach, et al., 1972). Validity studies then examine intended use situations, which typically vary more widely, are more complex, and are not under the control of the examiner (Kane, 2006). Although context is in the background in the measurement tradition, analyses by practitioners and researchers to address both practical and theoretical problems address considerations that are important from an interactionist perspective; see for example Messick (1994) on design and validity in performance assessment and Bachman and Palmer's (1996) and Davidson and Lynch's (2002) analyses of task features.

## Context

Task and test specifications address context to the extent that tasks exhibit relevant features, but they background the capabilities of persons to act in situations. L/C/S models connect the two, and connect both to data concerning performances. *External* context refers to features of a situation as they are viewed from the outside by the assessor, through the lens of the relevant L/C/S models. The focus is on the L/C/S models in terms of which claims about examinee's capabilities are desired; that is, the features of the situation (some of which the assessment designer may have arranged, others of which arise through the examinee's interactions in the situation) that should activate appropriate actions on the part of the examinee.

Whether an observer will see such actions depends on the examinee's *internal* context, or the situation as she views it through the mental space she constructs (Wertsch,

1994). The examinee's actions are evaluated in terms of the targeted L/C/S models (e.g., appropriateness, effectiveness, fluency, correctness). The internal context the examinee constructed may or may not have incorporated the features that the assessor deems relevant, in terms of the targeted L/C/S models. Even when the examinee's situation model incorporates the salient features in targeted ways, the examinee might not act effectively or appropriately. Observable actions are imperfect indicators of the model that an examinee constructs in the task situation, and are even more fallible as evidence about the situation model that she might construct in a situation with different features.

Standard test analyses speak of examinees' capabilities in terms of how well they perform on some set of tasks, and task difficulties in terms of how well examinees do on them. These concepts serve well for sets of homogeneous context-lean tasks. Structured item response theory (IRT) analyses in these contexts add insight into the nature of proficiency by examining the features of tasks that make them difficult ("difficulty drivers") and the corresponding nature of capabilities that examinees are able to bring to bear on them. Such analyses examine task difficulties in terms of cognitively-relevant features such as working memory load, number of steps required, and complexity of syntactic structures (Leighton & Gierl, 2007). This work exploits what Robinson's (2001) calls "complexity factors" for reasoning in terms of a targeted set of L/C/S models: What features drive difficulty for examinees who act through these models?

Unidimensional analyses are less satisfactory for context-rich tasks that require a greater range of L/C/S models in coordinated use (Mislevy, Steinberg, & Almond, 2002). Tasks can differ in difficulty in the traditional sense (e.g., percents-correct) for different reasons, but what makes a task difficult for one examinee may not be the same as what

makes it difficult for another. Beyond complexity factors, another factor contributing to difficulty is whether a given examinee happens to have experience with and can activate relevant knowledge and actions. Robinson (2001) calls these "difficulty factors." What makes medical knowledge tasks in German difficult for German medical students is the medical knowledge, but what makes them difficult for American doctors studying German is the German language. This variation, systematic though it is, constitutes noise or error through the lens of a unidimensional proficiency model.

Two ways that assessors can deal with complex tasks are design and analysis (Mislevy, Steinberg, & Almond, 2002). Design strategies call for principled patterning of features in tasks. Analysis strategies use multivariate models to sort out different sources of difficulty among tasks and correspondingly different profiles of capabilities among examinees. We reiterate that exhaustive analysis of all features of task situations and all aspects of capabilities in terms of L/C/S models is not possible. The features to model are the ones that are most pertinent to the L/C/S models that are focus of claims. We see next that many features of models need not be modeled in practice when task situations, use situations, and examinee characteristics are matched appropriately.

## Conditional Inference

What makes medical knowledge tasks in German difficult is different for German medical students than for American doctors learning German, even when both groups face identical tasks with identical constellations of features. But if the assessor knows that all of the examinees are American doctors studying German, she can rule out the alternative explanation of poor performance being caused by an examinee's lack of

medical knowledge. Performance then grounds claims about these examinees in terms of German proficiency conditional on medical knowledge. If the assessor knows that examinees in another group are German medical students, the same tasks can ground claims about medical knowledge conditional on German proficiency. This simple example offers important insights:

- An assessor's prior knowledge about the relationship between the task demands and the capabilities of an examinee plays a role in an assessment argument.
- This information is not part of the task situations per se, but a condition of the assessor's state of knowledge. Assessors with different knowledge (teachers versus state school officers), would be able to draw different inferences from the same data and would be vulnerable to different alternative explanations.
- This knowledge does not appear in analytic models for assessment data, such as student variables in measurement models, yet it is integral to the variables' situated meanings.
- Tasks alone do not determine "what a test measures," but also the assessor's prior knowledge of capabilities examinees are able bring to bear in what kinds of circumstances, on which they can condition inference.

The inferential burden of the examiner using complex tasks is simplified when it is known or can be arranged that certain demands of tasks, in terms of L/C/S models that are not the primary focus, are within the capabilities of an examinee. Even though the tasks may be demanding with respect to these capabilities, the examiner's knowledge allows inference to be conditional on their possession. Measurement models need have

not student variables for these capabilities, although the meaning of the variables that *are* in the model is now conditional on them. Configurations 3 and 4, discussed below, take fuller advantage of conditional inference than Configuration 2.

**Configuration 2, continued: Lean context, predetermined targets, for broad purposes**

Consider again the configuration that produces, for example, language tests called reading, writing, speaking, and listening proficiency. Examinees use targeted language structures and strategies to achieve limited task goals, in lean contexts and with little interaction. The validity of the assessment use argument concerns the extent to which an examinee might bring language similar structures and strategies to bear in TLU situations. We can analyze this situation with Messick's (1994) conceptions of construct under-representation and construct-irrelevant variance, developed under a trait perspective but now viewed in terms of L/C/S models.

*Construct under-representation* occurs when 'the test is too narrow and fails to include important dimensions or facets of the construct' (Messick, 1989, p.34). This description comports with recognizing that task situations do not involve key features of TLU situations that would require a person to activate and act through relevant L/C/S models. That a person can activate and interact through appropriate models in the task situation does not guarantee that she will do so in the TLU situation. Conversely, decontextualization and abstraction in the task situation can fail to provide the cues that would activate targeted capabilities in the TLU situation. People struggle with proportional reasoning test items yet carry out procedures that embody the same structures when they buy groceries (Lave, 1988). This is unsurprising from an

interactionist perspective. It constitutes an alternative explanation in an assessment argument for success in a criterion TLU situation despite failure in simpler task situations that are formally equivalent.

*Construct-irrelevant sources of variance* arise from features of the task situation that are not reflected in the criterion situation. They can render success on tasks more likely or less likely than in the criterion situation, due again to mismatches in constellations of features across situations and the L/C/S models the tasks activate in examinees. Task features that make success more likely include familiarity with formats, cues, and social contextualization. Task features that make success less likely include demands for specialized L/C/S models not needed in criterion situations (e.g., a computer interface), and lack of context, motivation, interaction, meaningfulness, or feedback that a criterion situation would have.

The way forward, Chalhoub-Deville (2003) asserts, is examining which patterns of behavior are robust, both across and within persons. Which test situations call for L/C/S models and evoke performances that are likely to be similarly evoked in what kinds of criterion situations by which examinees? How does variation in criterion performance for a given individual vary with what kinds of task features (generalizability), and what kinds of criterion features (validity)?

**Configuration 3: Rich context, predetermined targets, for focused purposes**

Tests of language for special purposes (LSP; Douglas, 1998, 2000) start from the desire to predict use of language in specific TLU situations: “If we want to know how well individuals can use language in specific contexts of use, we will require a measure

that takes into account both their language knowledge and their background knowledge." (Douglas, 1998, p. 282). More is thus known about the features of the TLU situations in terms of required patterns of interaction, genres and conventions, sociocultural considerations, and substantive knowledge. LSP tests intentionally include features that mirror those of TLU situations, to require the joint application of L/C/S models that are required in those situations. When language rules are addressed in these tests, they are in fact interlocked with a language user's knowledge of when, where, and with whom to use them, to paraphrase Rod Ellis (1985, p. 77).

As an example, the University of Maryland uses the International Teaching Assistant Examination (ITAE; Maryland English Institute, 2008) to determine whether non-native English speaking graduate students can carry out the duties of a graduate teaching assistant. It includes an oral interview, a dictation listening test, and a ten-minute "microteaching presentation." While the interview and dictation test are similar to broad proficiency tests, the situations and topics concern the university teaching context and language use in that context. For the micro-teaching presentation, "the examinee explains a principle or a central concept in the field in which s/he is most likely to teach. … The examinee selects the topic. The presentation should be delivered as though to a class of undergraduate students who have no substantial knowledge of the concept. In the final 2 - 3 minutes of the allotted time, evaluators and others ask questions related to the presentation" (Maryland English Institute, 2008).

LSP tests employ features that emulate the situations and purposes with the aim of evoking the knowledge and interaction patterns of TLUs. In Toulmin's argument terminology, alternative explanations for both good and poor performance caused by

mismatches of L/C/S demands are reduced and the argument from observed performance in the test situation to likely performance in the criterion situation is strengthened. In Messick's validity terminology, invalidity due to construct under-representation is reduced. In interactionist terminology, the stronger similarities between the task situations and criterion situations make it more likely that an examinee will activate and act through targeted L/C/S models in both situations, or will not do so in both situations.

This configuration illustrates a point concerning conditional inference. An LSP test requires capability to use language in a specified space of contexts, and demonstrating that capability jointly requires knowledge of substance, practices, conventions, and modes of interaction in those contexts. Deficiencies in any of these areas can lead to poor performance. The nature of the claims that can be made about an examinee from performance in LSP tests depend materially on the assessor's information about the examinee's capabilities with regard to these considerations—specifically, about the examinee's familiarity, through experience, with such contexts in respects other than the focal capabilities. For example, when the ITEA is used to qualify non-native English speakers as graduate assistants, all of the examinees are known to have expertise in their substantive area. They are familiar with the contexts, mores, genres, and interactional patterns of academic situations, at least in the language in which they studied, including lectures of the type they are asked to present in their microteaching sessions. Choice of topic eliminates lack of substantive knowledge as an alternative explanation of poor performance. The claim is thus about using English to produce and deliver familiar academic material in a familiar situation. Both the assessment situation and the TLU

situation are complex, but the examinees' capabilities with many of essential aspects of the situations are known a priori to the assessor and matched in the test design.

**Configuration 4: Rich context, opportunistic targets, to support learning.**

The final assessment configuration is using context-rich tasks in instructional settings, to support students' learning with individualized feedback. As an example we consider task situations that are essentially equivalent to ITEA microteaching tasks, except employed instructional sessions that help examinees who did not pass prepare for a re-take. The same task situations, the same performances, and the same matchups among examinees' histories, test situations, and criterion situations remain in place.

Because the purposes and the claims they entail differ, however, the assessment arguments differ. They differ in a way that can be understood in terms of examinees' capabilities to instantiate and act through the L/C/S models that are appropriate to the task and criterion situations. When the microteaching task is used to screen graduate teaching assistants, only overall performance is reported. Because content expertise and situational familiarity were known a priori, it can be presumed that improving performance is not be a matter of, say, learning more about physics or criminology, or the semi-formal and minimally interactive lecture format. Rather, the sources of poor performance could include incomprehensibility, particular distracting lexical or syntactic problems, or an inability to marshal resources in the lecture context. Any of these by themselves or in combination produce a failing ITEA score.

*"Opportunistic targets"* means that the claims of the assessment argument are constructed in response to each examinee's performance. What specifically leads to

unsuccessful communication for this particular examinee in this particular performance? Are there problems with prosody or pacing? Are there distracting use of tenses, such as "John eated the apple yesterday"? The examinee's particular difficulties in this performance determine feedback and instruction for that examinee. The student's attention will be called to those aspects of English usage, as it is used in these situations and how it helps shape meaning in them, where it appears that targeted experience will improve her capabilities. As such, this assessment configuration and use exemplifies the "focus on form" instructional strategy in second language acquisition (Long, 1988).

## Conclusion

If language is a complex adaptive system, what is language assessment? Language assessment is gathering observable evidence to ground inferences about individuals' capabilities to use language in real-world, interactive, constructive situations, where the elements of the assessment argument are conceived, related to one another, and purposefully designed. This article sketches a way of thinking about language testing that is compatible with research on both language and on assessment. It provides a finer-grained and more situated understanding of the models and methods of the educational/psychological measurement tradition that guide practical applications. To flesh out the sketch, more detailed explication is required to connect the higher-level, more coarsely-grained narrative space of educational and psychological assessment with the lower-level, finer-grained narrative space of the interactionist perspective. Worked-through examples with specific tests, in terms of specific linguistic / cultural / substantive models, are needed.

# References

Bachman, L.F. (2003). Building and supporting a case for test use. *Language Assessment Quarterly, 2,* 1-34.

Bachman, L.F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, & D. Bayliss, Eds. (pp. 41-71). *Language testing reconsidered.* Ottawa: University of Ottawa Press.

Bachman, L., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Beckner, C., Blythe, R., Bybee, J., Christiansen, M.H., Croft, W., Ellis, N.C., Holland, J., Ke, J., Larsen-Freeman, D., & Schoenemann, T. (in press). Language is a complex adaptive system. *Language Learning.*

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 20,* 369–383.

Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32-70). New York: Cambridge University Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Davidson, F., & Lynch, B.K. (2002). *Testcraft: a teacher's guide to writing and using language test specifications*. Newhaven, CT: Yale University Press.

Douglas, D. (1998). Testing methods and context-based second language research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 141 -155). New York: Cambridge University Press.

Douglas, D. (2000). *Assessing language for specific purposes.* Cambridge: Cambridge University Press

Ellis, R. (1985). *Understanding second language acquisition*. Oxford: Oxford University Press.

Fauconnier, G., & Turner, M. (2002). *The way we think.* New York: Basic Books.

Gee, J. P. (1992). *The social mind: Language, ideology, and social practice*. New York: Bergin & Garvey.

Greeno, J.G. (1998). The situativity of knowing, learning, and research. *American Psychologist, 53*, 5–26.

Kane, M. (2006). Validation. In R. J. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 18-64). Westport, CT: Praeger.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.

Lave, J. (1988). *Cognition in practice.* New York: Cambridge University Press.

Leighton, J.P. & Gierl, M. J. (Eds.) (2007). *Cognitive diagnostic assessment: Theories and applications.* Cambridge: Cambridge University Press.

Long, M. H. (1988). Instructed interlanguage development. In L. M. Beebe (ed.), *Issues in second language acquisition*: *Multiple perspectives* (pp. 115-141). Cambridge, MA: Newbury House/Harper and Row.

Maryland English Institute (2008). ITA Evaluation. College Park, MD: Maryland English Institute, University of Maryland. Downloaded July 2, 2008, from http://international.umd.edu/mei/572.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement (3rd Ed.)* (pp. 13-103). New York: American Council on Education/Macmillan.

Messick, S.J. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Mislevy, R.J. (2003). Substance and structure in assessment arguments. *Law, Probability, and Risk, 2*, 237-258.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 257–305). Westport, CT: American Council on Education/Praeger Publishers.

Mislevy, R.J., Steinberg, L.S., & Almond, R.A. (2002). Design and analysis in task-based language assessment. *Language Assessment, 19,* 477-496.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3-67.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 22*, 27-57.

Strauss, C., & Quinn, N. (1997). *A cognitive theory of cultural meaning*. Cambridge: Cambridge University Press.

Toulmin, S.E. (1958). *The uses of argument.* Cambridge: Cambridge University Press.

Wertsch, J.V. (1994). The primacy of mediated action in sociocultural studies. *Mind, Culture, and Activity, 1*(4), 202–208.

Young, R. F. (2000, March). *Interactional competence: Challenges for validity*. Paper presented at the annual meeting of the American Association for Applied Linguistics and the Language Testing Research Colloquium, Vancouver, British Columbia, Canada.

Young, R. F. (2008). *Language and interaction: An advanced resource book*. New York & Abingdon, UK: Routledge.

List of Figures

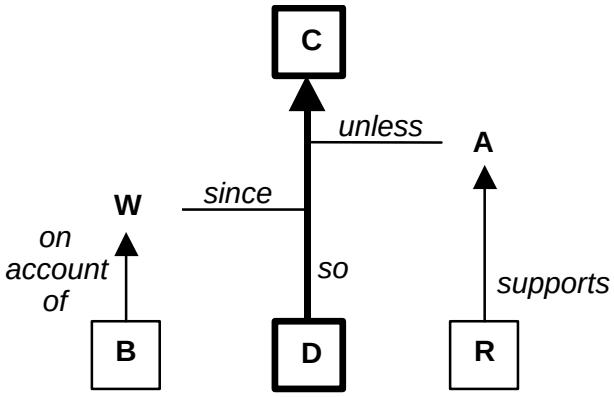


**Figure 1: Toulmin's general structure for arguments.**

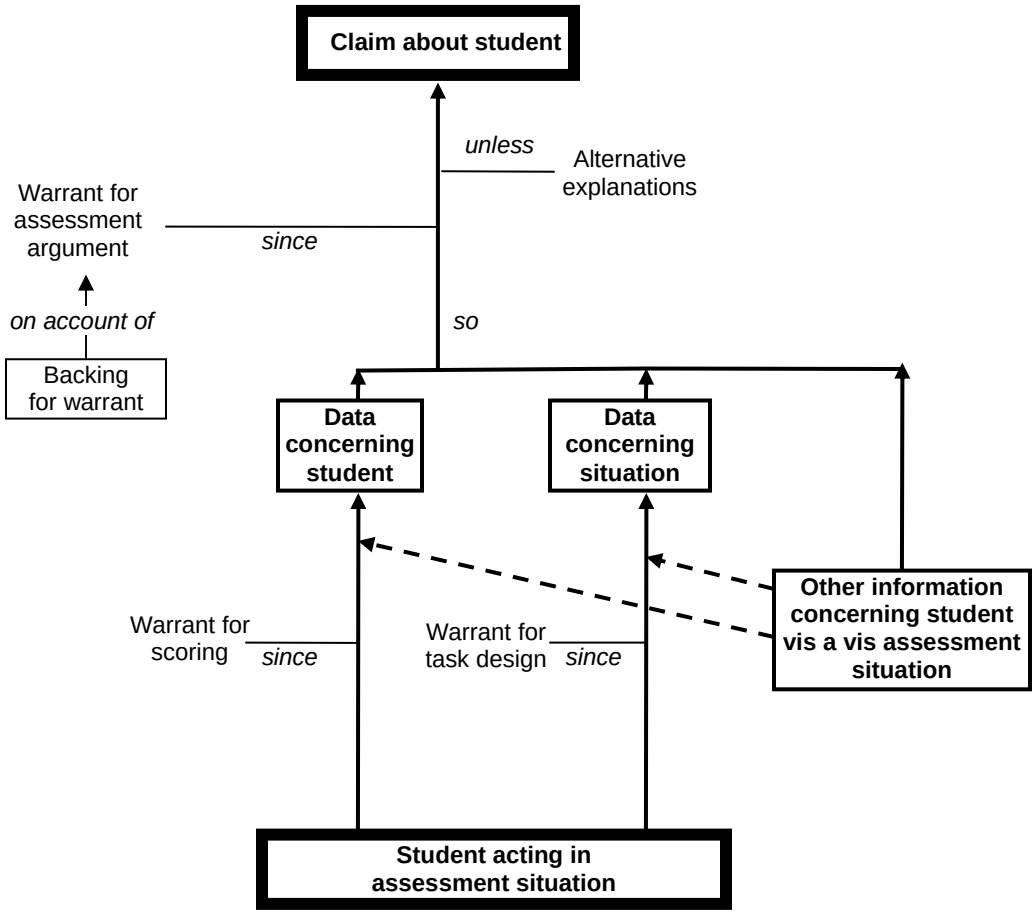


**Figure 2: The structure of assessment arguments.**

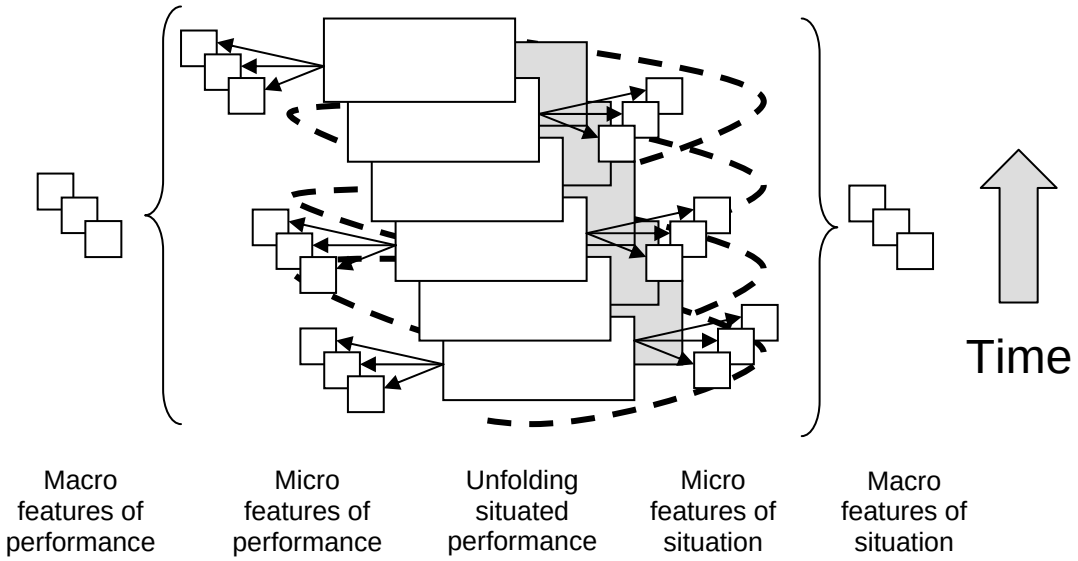


**Figure 3: Elaborated performance and data potion of assessment argument.**

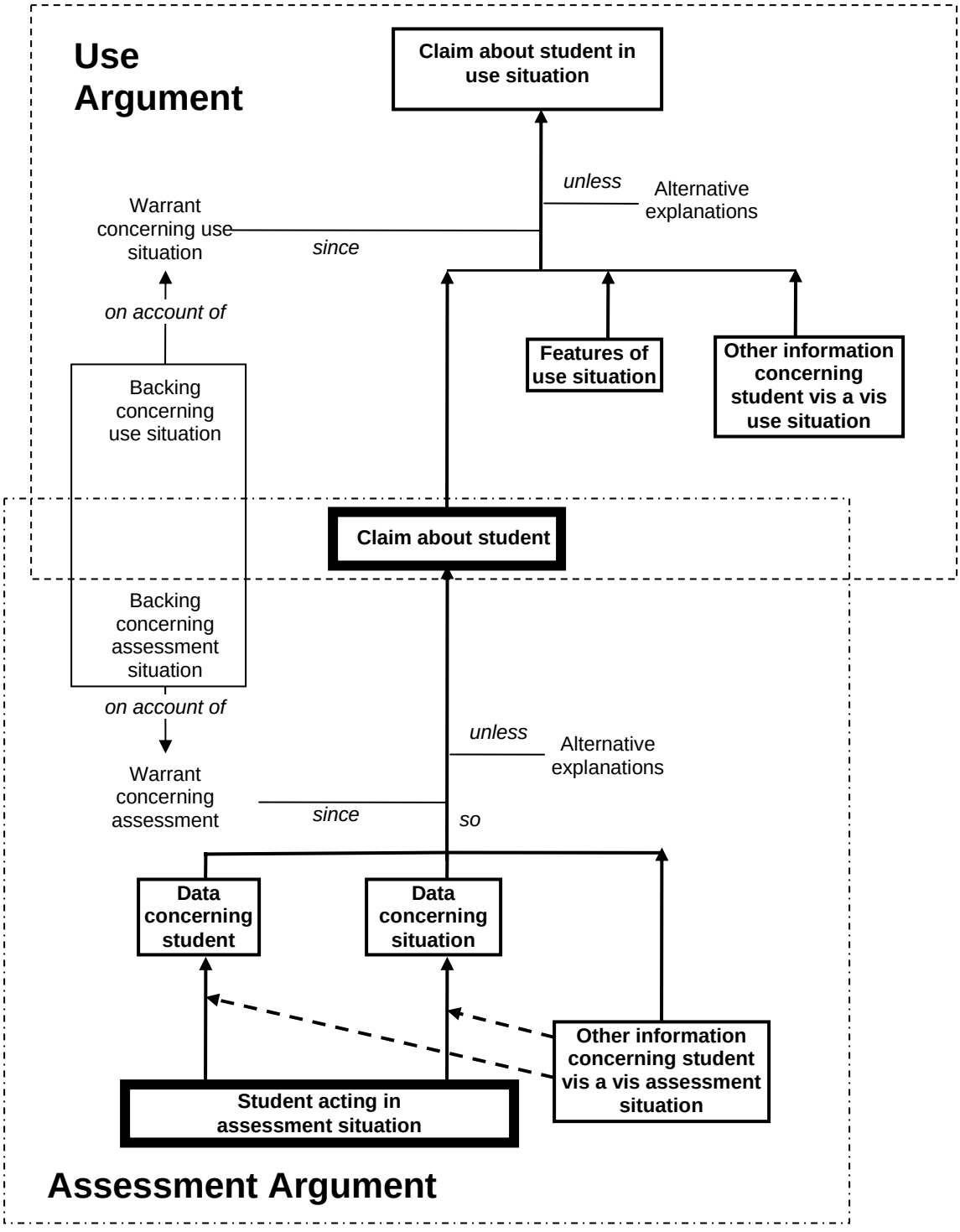


**Figure 4: Toulmin diagram for assessment argument and assessment use argument combined.**