Well, thank you very much, Freddy and David and Catherine and all the other folks at the new Center for inviting me.

As you notice, I kept assessment out of the title, another lure to come back after dinner and not bore you with that. But I've been thinking about all that we've heard and learned already today. And if you fall asleep after the first five minutes, I'm just going to tell you now the three crucial things that I think today's information boils down to for academic language assessment. So after I've told you these three things, you can tune out if you need to!

The first thing: I think we need to know what academic language is and how it develops over time. This is something that Guadalupe brought home to us very forcefully this morning—the need for longitudinal studies of what academic language is and how it develops—and we need this in order to be able to assess its progress and its ultimate attainment.

The second thing: creating an academic language assessment that doesn't inadvertently measure content area knowledge, such as science and math. It's still a messy art although we're trying to make it a science, as you'll see in a moment. And I think Debbie and Jana, this afternoon, brought home that issue of a fine line between when we stop measuring academic language and when we cross over into inadvertently measuring science or math and so on.

And the third thing: we cannot or we should not devise assessments of academic language separate from how we teach academic language in the content areas. Now, all the issues that I'm going to talk about in terms of psychometrics, the test development process and so on, those are just the tools of our trade. Those are the tools of the trade of the language tester and those are just for making a sound valid assessment, although I think what counts is especially that last point, point three. We need a coherent and integrated approach to instruction and assessment if we really want ultimately for assessments to contribute to a validity argument. The validity argument is that a test measures what it claims to measure and that it's meaningful and useful—useful to teachers teaching students, useful to a district reporting to parents and the state, and useful to a state reporting to the federal government. Okay, if you fall asleep now, it's fine!

Presented October 1, 2007; Oak Brook, IL create@cal.org www.cal.org/create

I'm just going to give you an overview of the presentation. I'm going to be talking about some background of this whole endeavor in terms of measuring academic language and a little bit of terminology so we share some common ground. Then I want to define some features of social and academic English and the implications that has for assessment. I will then describe the test development process that we have developed at the Center for the Study of Evaluation; if there's time, maybe share an illustrative prototype that we've created; and, finally, give some next steps, because we are a long way from having achieved our goal of a valid assessment of academic English.

So, actually, where I wanted to start was with a couple of personal interactions with academic language that I think are reasonably telling. These are both recent anecdotes. One was an e-mail that I received from a colleague at CRESST that very simply said, "Alison, what is academic language?" And I replied with a kind of defensive two-page response because it's not an easy question to answer. But in the process of thinking about my response, it really made me appreciate—perhaps for the first time—that academic language is really an interaction between a student and language. And by that, I mean, that each student has their own personal history or set of experiences with academic language. And so academic language is actually a moving target and it's very difficult to define, specify, and consequently, measure. So that was the first anecdote where I am personally studying academic language.

The other: It was very disconcerting to see as we're in the throws of trying to answer that question ("What is academic language?") that there was a recent TESOL Quarterly article entitled "Is there an academic language?" So the ground has shifted. Okay. That article by Ken Hyland and Polly Tse was questioning the validity, or the utility, rather, of academic language—the notion of a general academic language—because academic language seems to be so subject- or content-area specific, especially when you get to the higher grades and into college and so on. They really wondered whether there was enough language that cuts across the content areas to even claim there was such a thing as general academic language.

And, as you will see, the notion of academic language meaning something different for each and every student and the notion of general versus subject-specific or content-specific academic language are the themes that cut through the presentation.

But first, I want to share some definitions because these things have meant different things to different people at different times. You've had a chance to look them over right now. I'll just point out though *modality* – we're still talking about listening, speaking, reading, and writing as *modalities* although the Title III literature, I think, has shifted to the use of the word *domain*. But I will be talking about *modality*. And then *prototype*: again, it can mean something different to different people, and for our work, it's been an example task that is tried out and analyzed to established measurement for a specific construct (in this case, academic language) with a target population (for us, English learners).

So the background context – I know David actually covered some of this in his presentation as well as the ramifications for accommodations, and so I'll just say something specifically in terms of ramifications for test development – English language test development.

So, first of all, the educational need in the U.S. is for language tests that monitor ELD progress and the attainment of English language proficiency. Those of you who have been around for a few years may have noticed what I notice: this shift from ELD to ELP. Personally, the way I make sense of that, I think of the ELP construct as being this focus on reaching a level of proficiency analogous to being proficient at a proficient level in math or science and ELD as the actual development that takes place *en route* to that proficiency. Existing language tests are not good predictors of performance on standardized content tests, as David mentioned, and some of our own research has shown that there's a pretty big mismatch between the language tested and language being used on content tests and in the classroom. And so that led to this need for development of tests that measure academic English.

And the CRESST work that I will mention is really focusing on this age group that the new center CREATE has identified as a high need time in students' development for us to really understand what's going on.

Now, this figure has morphed a lot over the last couple of years. I think serendipitously the U.S. Department of Education really provided us an opportunity to define the construct "academic English" when they decided that states should show linkage between their Title I content area standards and their Title III ELD standards or ELP standards. And so that relationship, which is labeled with the arrow "A", in my opinion, is really an operationalization of academic English. It's the language that is inherent on those

content standards that we really should be trying to articulate and have represented in our ELD standards. And then those standards, obviously, can serve as a blueprint or a guide to the ELD items on the ELD test.

There's other elements there, obviously, that I could talk about but those – the link between "A" and then from "B" back to the ELD standards are, I think, the crux of the test development process and effort that we need to make.

Now, a lot of different researchers with a long history that people have mentioned today, are working in the area of academic English. All of these studies and more—this is not an exhaustive list by any means—have impacted the research that we've been doing over the last few years. And various aspects of the academic language construct can be synthesized and considerations taken from this research. And I've synthesized that into key features, which I'll show you in a moment. (And I'll just point out for those of you who were at the LEP Framework hearings in DC, this is the origins or the underpinnings of the remarks that I made at that LEP Framework meeting. I think Diane was there, a few other folks were there as well). So let me just mention then how I pulled the information from these different studies, including our own work, to come up with seven key features. And all of these—*purpose*, *formality*, *context of use*, *context of acquisition*, *modality* or *domain*, *teacher expectations*, and *grade level expectations*—all of these pose challenges for creating an academic English test.

There was just one handout for this presentation, and that is rendered into this slide. Those of you who still have your folders, you'll find it; you can peruse that at your leisure. I want to point out just a couple of things in the interest of time. One thing I want to point out is that these notions of social language; of school navigational language, for want of a better term; and curriculum content language don't necessarily correspond to the traditional notions of everyday language versus academic language. And by that, what I mean is that everyday language could potentially cut across both the school language as well as the school navigational language and academic language could cut across both navigational language, the language used to negotiate classroom management and things like that, as well as the curriculum content language. So that's one thing I wanted to point out to you.

This is the point that I wanted to re-enforce. I believe Guadalupe made this point, others this afternoon made this point, that academic language doesn't just mean written language. And I think David, in your presentation, when you talked about reading and writing being more of a predictor and you referred to reading and writing as academic language, I think I would really caution us, and I think others would to some degree as well, that oral language can also be quite academic. And I wonder actually in the case of your findings whether or not one explanation that you may have considered was that the medium that the kids were being tested in, it wasn't oral math. You know, they weren't talking about mathematics and that maybe there's a closer correlation because they were reading and writing about mathematics and science.

So I don't want us to forget then that modality is important and that at the higher grades, especially, I've heard teachers talk about academic language as if it were just the printed language—reading and writing skills—but oral language is still going to be critical. It's going to be critical for taking lecture notes so the listening skills are going to be important, a lot of oral presentations that kids do, discussion in classes. We've heard today from Diane and Debbie and Jana about having kids really communicate—display their academic knowledge through oral language.

There's a couple of different assessments (the state assessments, the consortia that have developed assessments for states) that also take a slightly different stance on some of these things, so maybe it's worth me just mentioning a couple. For instance, the ELDA, the *English Language Development Assessment*, put together by the CCSSO consortium: they don't appear in their construct definition to have social language assessed on that particular assessment. The focus is on the social environment of school which would be equivalent to the school navigational language that we have up there and also curriculum content language so that's conveying curriculum-based content in the language of the ELDA assessment. Another major assessment which was mentioned today, the WIDA, has in one area both social and instructional language which would be equivalent to the social and school navigational language on this chart as well as individual content area language associated with individual content areas: so English, language arts, math, science, and social studies.

The *Stanford English Language Proficiency* test, the SELP, which some states, I think, have taken and modified for their own purposes: Now, that particular assessment doesn't have any reading and writing modality for social language so reading and writing is only being assessed equivalent to the school navigational and the curriculum content dimensions on this chart.

So it's interesting, actually, to go look at the different assessments which are now available and see what constructs they actually cover when you break it down like this.

So implications for assessment: It is interactive and context dependent. As I was mentioning, it's different for each and every child. When you look at academic language like this, you see that it's dynamic and process oriented; it's ever-changing. English language development does develop. I think this notion of a differentiated assessment, the way we differentiate instruction, might be useful and falls out of that framework that I just showed you. You can imagine at least in classroom-based assessment teachers being able to individualize and target the kinds of academic language that they need to assess. And I think it's also suited, therefore, for computer-assisted testing.

There are also implications for test validity. There's a lot of variation in that chart in terms of children's opportunity to learn academic language. I'll say something at the end of this time about some new work that we're doing but to think of opportunities to learn, not just content, but opportunity to learn academic language and so that does have implications then for how valid your test is going to be if kids have not had the opportunity to acquire academic English.

And the next comment there about establishing native English speaker norms. We don't have enough information about how these different languages—social language, academic language—develop with monolingual English speakers, native English speakers, and we do need to know what those norms are so that we don't develop tests which are more challenging for our English learners than had we developed those tests for our native English speakers. And you'll see in our own test development efforts, we do include English only students all the way through so we know how to calibrate the test items and so on and know how others students would be performing.

One thing to point out: in the state that I'm living in, and so is Freddy, we are not allowed by law to give the CELDT, the *California English Language Development Test*, to native English speakers. So

that test in California, apart from some very early work, apparently, has never been (on a wide-scale, certainly, in its most recent forms) normed with English only students. And I think that's quite problematic.

And then this last main bullet. I think that if we can create a common academic English construct and establish those learning progressions for ELD—those longitudinal trajectories of how English language development develops— then we can have some continuity between formative assessment, benchmark assessments, and summative or large-scale assessments. Right now, these things seem to get developed in isolation. But if we can go to this underlying construct, well-articulated, we might be able to pull something off that's much more systematic an approach to assessment across the board and, I think, through that mechanism, we could forge a closer tie to instructional practice rather than look at these things in tandem. I think we are getting a lot of attention and funding for assessment, but I think it's not going to be very efficient or, at least, it won't be very cohesive unless we pay as much attention to the instructional elements as to the assessment.

So in terms of the Academic English Language Proficiency, AELP, the prototype developments at CRESST, this is the schema of the process that we used. This is the one we used but we have learned a few lessons and we would actually modify this. And, in fact, the report that you see there, which is available on the CRESST website, shows you the final rendition of the test development process that we would advocate. Basically, what it would involve is making this an iterative process. Right now, we only have feedback loops for revising our tasks after we had been piloting and using some early pre-pilot results. We did have an audit trail. We adopted this from Fred Davidson's work. Fred, who worked on the WIDA project, also had this audit trail approach. So we were able to document all of our decision-making along these different points. But two adaptations would be to take this through the entire test development process and also to think about these feedback loops going further. The results of empirically testing the tasks would go back to specifications and also our ramifications for the construct that we first define and the kinds of empirical evidence for its existence that we would collect. So this whole thing would be much more interactive.

Now, those of you who are state level folks, or even district level folks with decisions about what assessments to purchase or sponsor in your district, I don't know how many of you would know that there

are professional standards for this kind of thing. The International Language Testers Association has been very active trying to codify and standardize the test development process so that just as other industries have standards and codes of conduct and so on, there are guides to whether or not you are purchasing a well-developed assessment. And this test development process (displayed here) would help give you at least a guide to making that judgment and making an evaluation of its qualities and hoped-for validity.

So basically then, this focus that we had with kids grades 4 through 6, our goals were to design and isolate specific language features and, in that respect, what we were doing was very similar to the WIDA consortium—that we wanted to be able to look across the different content areas and pick out some major linguistic features that we'd want to know that kids are acquiring. And so it was designed to help determine whether a student has sufficient antecedent knowledge of English language features to make meaning of academic text that they encounter, and just as Diane pointed out in the last session, that the authentic nature of this reading that we'd be asking students to do was paramount for us. And so we actually used the text from three major textbook publishers that pretty much covered most of mathematics and a lot of science and social studies in California anyway. So the text that we built our tasks around are all authentic grade-level reading that kids would be doing.

At stage one then in our work, we did collect evidence—we hoped good evidence—for what language demands were and where they were. And so we did text analyses, we did classroom observations, content reviews of standards just to be able to identify and give enough specificity for test development. And the outcome then were linguistic profiles in the area of math, science, and social studies. We were cowards about English language arts. It's very difficult, I think, in some cases to distinguish English language arts and English language development. They are separate content areas, but we wanted to, as cleanly as possible, make the case using these content areas where we wouldn't be too distracted with some of the more subtle differences between English language arts and English language development.

Stage two then was to integrate that empirical evidence. We took the texts selected from the published textbooks, as I mentioned, and created tasks using those linguistic profiles and the outcome

was, in the end, 101 different tasks that were created across the areas of math, science, and social studies.

At stage three, we broke things down into two phases, and you'll see in a moment we actually just finished with phase two. But at phase one, basically, we weren't prepared to call these pilots. They were too small. They really were for us to first figure out what was going on in terms of how long these tasks were taking kids to do and so on. And we also predominately focused on English proficient or even English-only students. For that reason I mentioned before, we didn't want to set the bar so high that a native English speaker would have had difficulty had we given them these tasks. So we did have an over-sampling initially of relatively proficient or native proficiency.

We needed more information than just the kids' test scores, of course, so we had demographic information, the teacher reported their reading level, and so on. And in the end, we got rid of about half of the tasks for various reasons: either they were not discriminating between the kids who were good readers and poor readers or there were complications with kids following the directions, you name it, but about 50 percent were gone. And that's typical: If you talk to folks developing language tests, they usually over-make the number of items they need because they anticipate that they're going to lose about half of them.

This is actually my favorite quote from the verbal protocols. We had a small number of 18 kids who we tape-recorded and sat with as they were completing the tasks. And so this is my favorite. This meant that the kids knew that this was not a math task. The danger was we would create items which were replicating real mathematics items and we didn't want to do that; we wanted to capture language. And so this was one of the youngest kids that we had in that pre-pilot sample who told us that, "You know, you didn't ask a big math question; it just asked you what she bought. It was telling you right there which." (You'll see this item in a moment and you'll know what that means). But the important thing was at least they recognized it really wasn't a math question.

Not all of them turned out to be that way; we did have to jettison them.

So phase two then. We did have far more students in terms of English language learners, and we collected some standardized assessments, not just teacher reports and so on, and we continued with that audit trail.

In the end, out of the original 101 items, we only really had 17 which we felt did the job of assessing academic English. [We removed the other 84] for all those reasons I said before: they may have been actually too close to the content area, requiring prior knowledge in science or social studies in order to be answered; they may have been badly written; a lot of test items are just bad items; the directions could have been confusing to the students. And so I think that's quite telling: if test developers—commercial test developers—tell you they throw away about 50 percent; well, maybe they're not throwing away enough. I don't think we did a particularly bad job here. My colleagues at CRESST included Carol Lord, Frances Butler, and Robin Stevens. These are people who have been around for some time in the area of language testing, so I think this is probably because we were catching very early on in the test development process items which were really failing to distinguish kids who were scoring highly on English language arts versus kids who were scoring poorly and so on. And with the process that we created, we were able to see this perhaps more clearly than some other test development efforts.

So just a quick illustration from our first pre-pilot version of some of these tasks. The task specifications were all based on standards for the State of California for both English language development and the content areas but were also informed by what kinds of tasks we were developing. Now, I'm using the word *informed* because we were careful not to make *standards-based* assessments. We were incorporating other information as well—the classroom observation, our linguistic analyses of different texts as well as tests—and so it's really not a standards-based approach; we're calling it a standards-informed approach. This is the typical test specification for a math-based language item. I'll just give you a moment to take a look at that.

So an example of an item that was built on these specifications would be the following. Now, I'm not particularly wedded to this item so if you have lots of criticisms of it, that's just fine. But basically, this was an authentic word problem in a textbook. "Carlotta bought nine packages of lemonade for $1.10 each

and two packages of cups" and on it goes. And then there's even the math question there. "How much more money did Carlotta earn than she spent on supplies?"

Well, we're not going to ask them to do the mathematics; they're not going to solve that word problem. Instead, we built a set of questions about the language used in that word problem. So this particular item, it wants to know basically what the word problem is getting at. In other words, don't do the mathematics; tell us what the problem should be. What is the word problem asking about? What is the concept, if you like—the knowledge that they need to bring to that? Is their understanding that the word "profit" would paraphrase what the math item is asking them to do?

And let me just show you here: This is the linguistic profile that we built. We had master profiles for what mathematics items' linguistic features looked like, and then we were able to use those master templates or profiles to build items that looked like or mirrored the overall master profile. So this is the specific profile that went with that item, and as we built that item, it had to closely mirror the features of the mathematics overall profile.

And there are a couple of examples here from the verbal protocol in addition to the one I showed you already about it not being a big math problem; "it just wanted to know how much she had spent" or whatever it was that the student had said. But these are a couple of other examples. So we have a student who is a 6th grader saying, "I didn't exactly understand. So I went back up the passage and read the questions that they asked. So then I noticed that *profit* is basically the same thing *earned of*—how much she'd earned. So it means how much profit; profit means the same thing." So it's a little disjointed. It's a little inelegant, if you like, but they were able to link the word "profit" in the choices and eliminate the other distractors.

Now, we had a bunch of statistics. These are the psychometrics I was telling you about. We calculated item difficulty and we wanted to know whether it would discriminate between kids who were performing well on a separate assessment of English language arts or if, in fact, it wasn't. So we had some basic psychometric information on the kid's performance on these items. And this item, we decided we would retain it and give a full description of it as a prototype of academic English language proficiency. And we have a number of those items, actually, in that report that I referred to earlier.

We also illustrate the ineffective items—the ones we had to throw away because they were not performing well (not the kids but the items weren't performing well). They were not discriminating good readers from poor readers. They were not discriminating based on home language background and other factors like that.

So there are limitations. This approach, you probably don't see it everywhere because it is immensely labor intensive. At the end of the day, all we can give you are 17 items that are built with basically the 5th grade as the target because with the 4th graders and the 6th graders, they were the calibration on either side. And so the 5th grade was the middle grade, the target grade, to find out how well kids were doing one grade lower and one grade higher. It does require extensive empirical analysis of the standards and the English language that's being used in classrooms.

At the classroom level though, we really recommend exploring tools, rubrics, specifications like the template that you saw for all teachers to identify academic language demands in the standards and other curriculum materials. People can do this for themselves. Once you have these profiles built, you could imagine being able to apply it to more than just a testing situation. You might want to take it to analyze your own textbooks and see what kinds of inherent language demands they contain. Or you could create your own classroom assessments—and I think of this in terms of you could do this in advance of and separate from the content assessment. So you may find out whether or not your students are doing well on math. And, on top of that, you could try and find out whether a student is doing well in the language of mathematics.

And for next steps, well, we could expand this to other grades. For example, there's a project at UCLA which is a collaboration with LAUSD, *Arts in the Middle Program*, and this AIM project is using this notion of academic English language proficiency. I guess these days it's not enough that you have an arts program; you have to show that there's other learning going on. And so the AIM project is actually linking its curriculum to English language proficiency for English learners, and so, in this case, it's visual and performing arts. And so we were able to take a look at the curriculum that the AIM project is using and extrapolate and work with teachers to identify what spoken language skills kids would need—what listening and speaking skills are being demanded of them when they're in a theater group, for instance.

And so we've been able to do a little bit of additional work in that context around the listening and speaking modalities.

And then, finally, additional CRESST studies that we're involved in are still trying to define what academic language is, and we've started to think of this as linguistic pedagogy. And in one particular study, we are working with a number of different states right now (and I think Barbara Medina is here from Colorado who we are working with), and we just put together a survey of academic language exposure. And, as I said before, we think of this as a part of this opportunity to learn construct. And in Aida's work—those of you who had a chance to look at the paper that she made available to us for the conference—she talks about macro and micro levels of scaffolding. So macro is planning and micro level scaffolding is the moment-to-moment scaffolding that we've heard about today. And so we're actually surveying teachers to find out whether – first of all, if they're aware that there's a difference between ELL strategies and academic English language strategies. In fact, is there a difference? I'm curious to find out whether other people would agree that you could distinguish ELL strategies from academic language strategies. But we're trying to survey teachers to better understand how much exposure they think they're giving their students for the opportunity to learn academic English. And then the last project with 6th grade math teachers is really just looking closely at how math teachers are explaining mathematical principles and concepts to their students. Again, all with this purpose of better identifying and articulating what academic language is so we can better assess it.