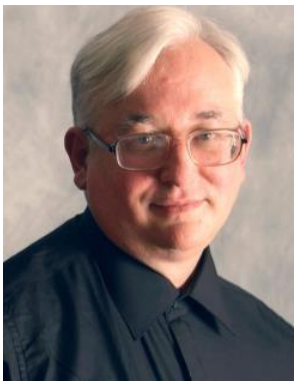
An Interview with APPLE Lecture Speaker Dr. Paul D. Deane

Peter Kim, Jorge Beltrán and Soo Hyoung Joo
Teachers College, Columbia University

INTRODUCTION



On February 25, 2022, the *Studies in Applied Linguistics and TESOL (SALT)* journal had the great pleasure of interviewing Dr. Paul D. Deane, the invited speaker for the 2022 APPLE Lectures Series hosted by the Applied Linguistics and TESOL Program at Teachers College, Columbia University. Dr. Deane was kind enough to take the time to speak about his research, his work on scenario-based assessment, automated scoring and his thoughts on cognition and L2 assessment.

Paul D. Deane is a principal research scientist in Research & Development at the Educational Testing Service. He earned a Ph.D. in linguistics at the University of Chicago in 1987. He is the author of *Grammar in Mind and Brain* (Mouton de Gruyter, 1994), a study of the interaction of cognitive structures in syntax and semantics, and taught linguistics at the University of Central Florida from 1986 to 1994. From 1994 to 2001, he worked in industrial natural language processing, where he focused on lexicon development, parser design, and semantic information retrieval. He joined Educational Testing Service in 2001. His current research interests include automated essay scoring, vocabulary assessment, and cognitive models of writing skill. During his career at ETS he has worked on a variety of natural language processing and assessment projects, including automated item generation, tools to support verbal test development, scoring of collocation errors, reading and vocabulary assessment, and automated essay scoring. His work currently focuses on the development and scoring of writing assessments, as well as examining use of keystroke logs to analyze writing in high-stakes assessments and in learning environments.

Looking Back

1. What got you interested in applied linguistics and specifically writing assessment? What have you enjoyed the most about doing writing assessment research?

I didn't start planning to get into writing assessment; I did my PhD at the University of Chicago on theoretical linguistics. The program in Chicago is extremely interdisciplinary. Taking courses in cognitive psychology, as well as linguistics and getting a broad-based interdisciplinary foundation with a heavy cognitive emphasis were what really prepared me for the writing assessment focus. I had an interest in writing because I've been thinking about where I was in teaching writing and was thinking about the implications of cognitive approaches to linguistics to writing and writing instruction, but that was not the main focus at first when I am joined ETS. A part of working with a very rich array of people at ETS is an opportunity to do more

cognitively oriented work. At ETS, I got involved in working with the CBAL® project (Cognitively Based Assessment *of, for,* and *as* Learning), and that's where I started moving more towards writing assessment. I was also a part of the NLP group at ETS supporting the automated scoring work and e-rater. I had developed a couple of the features that are in the rater engine, one focusing on syntactic variety and then another feature that focused on vocabulary richness. CBAL® involved every aspect of assessment design development from designing the construct to figuring out how to measure it and doing the automated NLP work to develop tools that would give you better insights into writing. At CBAL®, I ended up being the person in charge of writing assessment because that was the most natural fit to my areas of expertise. After a while, the reading and writing strands were merged and I headed the ELA team.

Scenario-Based Assessment

2. **Influence of the context.** You have worked on several projects from the CBAL K-12 system model, which "proposes not only to validly measure student competency but also to affect it positively by encouraging change in teaching and learning behavior" (Bennett & Deane, 2016). You have worked on the study of various aspects of CBAL® assessments over the last few years. Could you share with us, based on your experience, what have been the challenges of developing the contextual features of the CBAL K-12 system(s)?

I think the fundamental challenge in terms of the US system is that everything tends to be driven by the end of the year test. The whole concept of CBAL® was trying to go against that, so that we were thinking in terms of periodic assessments where the scenario-based assessments were given over the course of the school year. You are aggregating information, over time, rather than doing an end of the year test. In practice, we had to emphasize things that could be used effectively over more formative assessments. Everything we were doing was only really meaningful to the extent we could connect it to curriculum. With the scenario-based assessment, it is a compromise between two conflicting forces. On one side, you have the pressure for valid reliable measurement in the traditional sense with a three- or four-hour test at the end of the year as the way to make sure that you have validity and reliability. On the other side, you have the pressure to do something that gives you a measurement useful for diagnostic and formative purposes that is meaningfully integrated with instruction. And scenario-based assessment is in some sense a compromise between those two forces. It gives you a valid and reliable measurement. The correlation between our 90-minute reading and writing assessment and the Smarter Balanced assessments was almost as good as their correlation year to year. It was quite a reliable assessment, even though it was half the time. I think we did pretty well in the validity and reliability department, but one of the things I wish we could have done more of was do more integration with curriculum and instruction. The catch is to have digital tools that integrate with curriculum instruction. And we are talking with various teachers and colleges about reading and writing projects and about the ways in which assessments might link to curricular issues. We can hopefully influence curricula and assess things that are valuable. We've been told that looking at our assessments has taught them about things that they should have been teaching but weren't.

3. **SBA.** Now, I'd like to talk about scenario-based assessment. This technique is used for CBAL's summative component, among other reasons given that scenario-based performance assessments present "reasonably realistic contexts and formats for problem solving" while allowing for relatively independent measures of student competencies. This compromise, along with the opportunity to systematically model the social context has led to the examination of this type of design in L2 assessment. However, one concern has been that of generalizability and topic effect. In your 2016 study with Randy Bennet, the latter was found to be a relevant context for the measurement of argumentation skills (Bennett & Deane, 2016). What is the role of 'topic" in a scenario-based test? What assumptions would you say need to be evaluated in terms of topic as context?

One thing that's probably worth noting is that there are two different effects. One is the effect of the contextual framing and then there's the effect of topic. The effect of topic seems to be relatively small, in the sense that based on factor analyses, and you can get a factor of a topic effect, but it wasn't huge. On the other hand, topic effects are also related to a really critical issue, which is prior knowledge. And one thing we've found is that there is a threshold effect for knowledge. Given foundational knowledge, you need to be at a certain level before a scenario-based assessment really makes sense. If you are unable to learn from reading because your decoding skills are really weak or if you don't have enough prior knowledge to integrate what the text says with what you already know, then you're going to be very challenged and a scenario-based assessment may not give you the best information

At ETS we have what is now called Read Basics. It provides a series of assessments of foundational reading skills and the ability to use those skills in context. If these skills are critically weak, then the scenario-based assessment may not be the most appropriate test for those who need help in those particular areas. This implies that it is really useful in an instructional context to know how much people know about the text and whether they're prepared to understand the text. We're currently doing research based on the work from my vocabulary where we basically generate topical vocabulary assessments. We asked people to do a quick assessment of words that are related and not related to a topic. And what we found is that it is an efficient measure of prior knowledge that you could use to place people. Now context, of course, is ambiguous between topic and the general contextual framing. One of the interesting things that we found from some of our studies is that the contextual framing provided by the reading tests leading up to the writing task does have an effect on the people's writing but it's not the fixed effect you would have expected. We have conducted a study that varied the context by having the reading tasks before the essay or have the essay first and then do the reading tasks, or switch topics between the reading tasks to see what was going on. Not too surprisingly, switching topics in the middle of a task is a very bad thing. Interestingly, though, the major effect of the reading tasks on the writing was to decrease the cognitive load without increasing the quality. But the result was that the students who didn't have the preparation were working harder and producing more text without an increase in score. It suggests that the students who had the reading tasks were writing more efficiently, but maybe not trying any harder once they met their quality level, and they were they were willing to stop.

4. **Technology and Assessment.** Technological developments and technology-delivered assessments continue to diversify the types of tasks that can be used in L2 and L1 assessments. At times, trends show the different sides of a continuum. For example, in

contrast to the prioritization of *authenticity* through the use of scenarios to elicit higher order skills and model the social context of an assessment, there is a trend in the opposite direction where for-profit companies develop decontextualized L2 tests in order to enhance predictability of the responses that might be accurately rated/classified using machine learning. What do you think is next when it comes to development and score-interpretation of technology-delivered assessments?

The context of assessment is changing very rapidly, and one of the projects ETS is investing in is personalization. It's about providing feedback, supporting learning products and providing an assessment in the context that's not necessarily high stakes. ETS is also investing in TOEFL essentials, which is a lighter version of the TOEFL. One of the fundamental issues with a higher construct-oriented test is that the design work is a lot more effortful. We are trying to find ways to identify repeated contexts that's worth assessing and developing scenarios for. But at the end of the day, those are contexts that are probably also recurring in a formative classroom environment. If you could provide a portfolio which demonstrate that you had decent performance in English in a variety of different contexts, and it was collected during your English education, a college may find that a lot more interesting than one of these short tests. However, the short test is also really cheap and easy, and it will probably tell you about low-level skills that are just basic prerequisites. If you don't have the fluency, you don't have the production, and if you don't have the comprehension speed, all those things have to be at a basic level before the more contextualize scenario-based evaluations can differentiate the learners.

5. **Assessment of Argumentation.** While your research on L1 argumentation writing skills have focused on the US K-12 systems, a portion of the participants were ELLs. Do you have any suggestions regarding how to assess argumentation skills of second language learners?

Written argumentation requires supplying specific subsystems of language. The stance taking elements of the language are really critical. A lot of academic language related to stance taking that isn't purely argumentation are required for analysis. Argumentation is making very heavy demands on the syntax, vocabulary and discourse related to staking claims dealing with evidence, evaluating ideas and so on and so forth. These are specific targets that go beyond general linguistic fluency that you need to worry about if you're trying to assess written argumentation. This is related to the issue of discourse flexibility. Can people shift their style to one that's appropriate for a given genre or task? Some of the work we're doing on writing trade analysis is related to the work that goes to Biber in the 80s where you look at different dimensions of variation in texts. One of the issues that you think about is to what extent are people producing texts that match the expected characteristics for the given genre, in this case written argumentation. This is where a blind NLP approach is potentially problematic because we actually know pretty clearly what linguistic resources people need to master in order to deal with these. It's a very specific register of language that L1 speakers may not be particularly conversant with.

Cognition and Learning

6. Building on the previous question, your identification of the key literacy practices (e.g., build and share information, build and justify interpretation) have greatly contributed to the understanding of skills critical in the English Language Arts context, but I believe many of those practices are critical for language learners inside and outside of the US K12 context as well. Are there any particular considerations for teaching and assessing these key literacy practices for language learners?

To the extent that you start differentiating key practices, you're also differentiating genres of texts that are related to those key practices which ties into L2 language functions. For example, if you're looking at a narrative, certain kinds of functions are coming into play that are not the same as an argument text. If you're looking at text in a different domain like history you are changing the sampling of the linguistic resources that are at play. There is an alignment between the higher order critical thinking skills and analysis skills that people need to be able to apply. If you communicate meaningfully about the content of a narrative, you need to have a decent vocabulary for emotion, for social relationships, and you need to be able to have mastery of the syntactic construction to express theory-of-mind kind of relationships: what people think about, what other people think about, and how they feel about how other people feel at various levels. These are different than the resources you need to deal with argumentation. One of the issues you then run up against is that if you only look at learners' L2 capabilities in a specific context and don't look across a variety of context, you haven't actually looked at whether they have developed the skills to flexibly shift modes that are appropriate for the given tasks. There is a variety of genres that are linked to critical thinking and analysis tasks, and those tasks are linked to different applications of linguistic resources. Without making those connections and creating that flexibility, English education is going to produce people who seem fluent but are severely handicapped in contexts other than the ones they have been trained in.

7. You have defined learning progressions, or the roadmap for the development of a target skill. For content areas such as math or science, these trajectories tend to be more explicitly observable. However, as you pointed out in the 2019 article (Deane & Sparks, 2019) language is a skill-based domain, where those developmental trajectories are often complex, and many subskills are intricately interrelated with one another. Given this complexity, it seems there might have been challenges in designing learning materials or assessments to measure students' learning progression. What were the challenges, and how did you address those challenges?

It really comes down to needing to do a deep dive into the cognitive literature and understand what the moving parts are from a cognitive perspective. For example, in writing, the big issue is working memory and how skills like transcription or idea generation are competing for working memory with other processes like text production, monitoring, and revision.

One of the things that you then have to worry about in learning progression is that higher-level learning progression presupposes mastery and not necessarily next steps in a conceptual progression. A lot of the work we've done has been looking for dependencies that enable us to define these thresholds. For example, foundational skills are really easy to recognize. We have

publications from various people looking at different foundational skills. There was a recent article about assessing writing that examined how typing speed has a threshold relationship to writing performance. That is, if typing speed is above a certain threshold it doesn't really have a strong relationship to score, but once you get below that threshold score, typing speed has a very definite relationship to score. There's a similar type of relationship with decoding where below a certain threshold, decoding has a massive impact on reading comprehension but above the threshold, it is much less important; likewise, there is a similar relationship for prior knowledge. For a lot of these foundational skills, you can establish thresholds, but things get a bit more complex for more advanced skills. In argumentation, one of the tasks we have published a lot about is a simple classification task that classifies statements as for or against a position. This has a pretty strong prerequisite relationship to performance on the writing test in that people who are not accurate on that task do not generally produce more than a few sentences or short paragraphs. They don't produce the full essay and that is probably because if you are not able to handle the intuitive understanding regarding which side the argument is on, generating arguments is probably incredibly difficult for that learner. Interestingly, this particular skill also mediates the effects of decoding, prior knowledge or other foundational skills. In general, if you are low on foundational skills, you are likely to struggle on the simple classification task. But, having the foundational skills also does not guarantee that you'll do well on that task. You get much better prediction from the argument specific comprehension tasks, then you do from the generic foundational skills tasks.

8. Because L2 learners are influenced by multiple factors, such as their first language or when they first started learning the L2, many L2 teachers find it challenging to define the learning trajectory or what it means to have a mastery of a certain level. Even standards such as the Common European Framework of Reference for Languages (CEFR) have been criticized that they have not been based on the actual process of L2 acquisition (North, 2014). Do you have any suggestions for how to define L2 learning progressions, or how to modify/apply the literacy learning progressions in L2 teaching and learning contexts?

In vocabulary acquisition, it makes sense to talk about grade levels of vocabulary acquisition for L1 learners. This is because the vocabulary learners are exposed to as they move through a normal life experience is fairly predictable at a statistical level. However, it's not so clear in L2 context where a lot of vocabulary learning is intentional. The order of vocabulary acquisition may be strongly influenced by curriculum at that point. I think that is true about most of the linguistic elements in L2 acquisition. For example, if you're not exposed to texts that have the characteristic structures of literary argumentation, narrative or academic argumentation, including linguistic constructions and vocabulary, then there'll be a gap. And that is very much affected by the curriculum.

This comes back to a distinction that I learned about by accident on a trip to a conference a long time ago, where I met some teachers who came out of the French educational system in Canada. They drew a distinction between pedagogy focusing on how to teach a student versus how to teach a subject. And those are complimentary perspectives. The US educational system tends not to focus on how you present a subject well for a learner. If you are thinking about how to present content to a learner, there is a progression there, which is not a “learning progression” in the more pedagogical sense but an instructional progression where you are thinking about how

to order your material in ways that maximize flexibility and mastery. I think in an L2 context, this is even more critical. The key concept gives you a way to identify sets of related resources that need to be applied in the common context.

Keystroke Logs and Automated Writing Scoring

9. You've also done extensive research evaluating the L2 writing process through keystroke logs. Can you briefly explain how keystroke logs can provide information about the writing process or the proficiency of the writer?

Any kind of logging system captures information that goes beyond what you can measure from the final product. The process unfolds in real time to get a sense of what somebody's writing process looks like. For example, you have three people and each of them produce a short text. One learner writes very fluidly and easily and then stops. Another one spends 15 minutes rewriting the second sentence. And the third one takes a long time because they keep backtracking over individual words and having a lot of text production difficulties. These three people are different, and they represent different skill profiles. One advantage of eye tracking is that it helps disambiguate pauses and texts production. If somebody has a long pause, does it mean they are disengaged, they are reading a text, working on a piece of scratch paper to develop a plan, or does it mean they are stumped into just sitting there frowning and trying to get their themselves back together? There are a lot of different interpretations for a really long pauses and that's something that a keystroke log by itself does not tell you. What the keystroke logger really does is it gives you a sense both of low-level skills like typing fluency and the distribution and length of pauses, which give you information about other processes. Pauses at natural junctions such as sentence and paragraph boundaries or clause or other natural pause locations for sentencing, planning and idea generation tend to be more related to idea generation which is essentially the linguistic process of text production. Whereas if there are pauses and then jumps to different locations, the writer is not doing text production. They were probably reading their text and doing some kind of evaluation and acting based on the evaluation. Looking at the difference between jump edits versus pauses before burst of text production tells you something really important. For L2 writing, this is the notion of burst length: how many words can you produce before you have to pause to think. This is a pretty well-known measure of fluency of linguistic production, and it is something that a keystone log gives you a fairly direct measure of. Certain behaviors like hesitation and back spacing over your entire texts are surprisingly strongly related to performance on some of the short TOEFL junior writing tasks.

10. Wiggins (1994) once remarked that "all assessment is subjective" and relatedly, in automated scoring, human judgment is involved in identifying what aspects of the texts are most pertinent, i.e., feature engineering. Traditionally, essays were graded on three overarching features of language, content, and organization. What do you consider to be the most important features of writing for predicting/classifying proficiency?

Let's distinguish between describing proficiency versus projecting their score. You can get prediction of scores just from the number of words in the essay. Since almost all language

measures are related to fluency in some way shape or form, it is hard to avoid number of words having an influence on your features. At a very low level, if somebody is not fluent, they can't produce at least 150 words in an essay. But once you get above a certain length, then you start having differentiation and it is worth thinking about it in terms of a multi-dimensional space rather than thinking of it in terms of just one specific thing.

There are elements that are related to organizational development that are often related to length. There are also aspects of vocabulary sophistication and there are aspects of linguistics sophistication. There is the ability to avoid error in text production which is related to monitoring and revision capabilities. And of course, there are deeper level skills of idea generation and making your ideas actually work in the context. So, you can't measure everything with automated methods, especially the deeper content focused abilities. However, you can capture a lot of different dimensions and a lot of the work I've been doing recently is focused on developing a training model that reflects not just the dimensions that are relevant to score, but also to register in genre. Currently, the training model we are working on uses data from ETS criterion system with millions of students' responses. It has 12 dimensions, and they correspond to dimensions of overall text structure and genre/register. At the end of the day, you care about not just predicting their score but understanding whether they are on track for producing a text with appropriate features for their genre.

Looking Forward

11. To wrap up, I want to ask you about the proliferation of automated scoring and the role of teachers. Given the advancements in automated writing evaluation and scoring, will human scoring become obsolete in the foreseeable future? What should be the role played by the teacher in making assessments of their students' writing? How should automated tools be used to inform instruction in the classroom?

I would like to point out something that goes beyond automated scoring which is the recent development of deep learning methods that allow for automated text generation. There are tools out there that I've seen in which more advanced stuff like deep learning models where given a small snippet of information, the model generates texts for you. The text can sound quite sophisticated, but it generally does not make sense beyond a paragraph or so in terms of coherence. If you take one of these tools and you pick a topic from one of our scenario-based assessments and plug that in, you get a text that reflects what is out there on the web, in terms of content. Then the problem becomes knowing how to edit it into something that actually represents a decent and coherent thought. I think the bigger issue is what to do when these kinds of tools become available because students will use translation or text generation as a shortcut and not be necessarily developing the level of L2 writing skills. From a scoring point of view, that means that you have to start worrying about whether the texts are produced on their own. There are multiple classification problems and one of the classification problems is how good is a piece of writing and does it represent a good faith effort? Classifying something as a good faith effort is a very different classification problem then classifying something by its quality.

REFERENCES

Bennett, R., Deane, P., & van Rijn, P. (2016) From cognitive-domain theory to assessment practice, *Educational Psychologist, 51*(1), 82–107. doi: 10.1080/00461520.2016.1141683

Deane, P., & Sparks, J. R. (2019). Scenario-based formative assessment of key practices in the English language arts. In H. L. Andrade, R. E. Bennett, & G. J. Cizek (Eds.), *Handbook of Formative Assessment in the Disciplines* (1st ed., pp. 68–96). Routledge. https://doi.org/10.4324/9781315166933-4

North, B. (2014). Putting the Common European Framework of Reference to good use. *Language Teaching, 47*(2), 228–249. doi:10.1017/S0261444811000206

Wiggins, G. (1994). The constant danger of sacrificing validity to reliability: Making writing assessment serve writers. *Assessing Writing, 1*, 129–139.