Contextualised judgements: A comparison of the rating criteria used to judge oral presentations in higher education and speaking performances in the TOEFL iBT™

Ana Maria Ducasse
RMIT University, Melbourne, Australia

Annie Brown
ACER, Camberwell, Australia

This study investigates assessment within oral academic assessment tasks, specifically focusing on the criteria used by discipline specialists and comparing them with those used to assess performance on TOEFL iBT™ speaking tasks. Three pairs of tutors from three faculties took part in verbal report sessions where they watched, rated and discussed the performances of ten native and ten non-native students completing first-year university oral assessment tasks in their discipline. The verbal report sessions were audio recorded, transcribed and segmented into meaning-based units prior to thematic analysis.

The features which emerged from the analysis were compared with those described within the TOEFL speaking rubrics. The analysis found that while there were some broad similarities in the focus there were also marked differences. Two of the three TOEFL strands (delivery and content) were well-represented in the academic tasks assessments rubrics and tutor discussion. However, the quality of the non-native students' language was only of concern when it was perceived as interfering with the student's ability to communicate. An additional focus in the assessment of university tasks was the use of academic skills, prompts and aids, non-verbal communication and engagement with the audience.

Key words: TOEFL iBT™ speaking, speaking rubrics, oral assessment tasks, academic skills, non-verbal communication

Introduction

Academic assessment, particularly at the undergraduate level, has traditionally dealt with written texts. However, the recent focus on fostering 'generic skills' or 'graduate capabilities' has led to the use of an ever-broader range of learning and assessment tasks. Project-based teamwork tasks, portfolios, and 'multicomponent' tasks, which

require students to undertake a series of related tasks, are increasingly popular across a range of disciplines (Barrie et al., 2012; O'Sullivan et al., 2012, Sadler, 2009). So, for example, students might be required to work as a group to research and prepare a written project proposal, which they then integrate into a written report, and later share with classmates as an oral presentation. Such tasks are additionally seen as more authentic and relevant to the work context than more traditional academic tasks. The need to foster not only written, but also oral communication skills has led to the widespread and systematic integration of oral activities such as presentations into subject assessments within undergraduate courses (Bearman et al., 2016; Bowden, 2010; Joughin, 1998; Pereira et al., 2016).

The study reported here addresses the relevance of assessments on the speaking component of the TOEFL iBT™ as a predictor of performance on academic speaking tasks. The claim put forward for the use of TOEFL and other tests of academic language is that performance on the test is an indicator of test takers' readiness to cope with the linguistic demands of their course. Such claims must necessarily derive from the findings of a rigorous programme of validation research. In the TOEFL iBT™ validity argument, the extrapolation inference of its validity argument is based on the warrant that 'the construct of academic language proficiency as assessed by TOEFL accounts for the quality of linguistic performance in English-medium institutions of higher education' (Chapelle et al., 2008, p. 21).

The evidence in support of this warrant derives from a number of studies. One area of research has drawn on observational, survey and corpus data to address the congruence in the task demands of the TOEFL speaking tasks and real-world academic speaking tasks (Biber, 2006; Biber et al., 2004; Brooks and Swain, 2014; Brown and Ducasse, 2019; Brown et al., 2005; Cumming et al., 2004; Rosenfeld et al., 2001; Yi, 2012). Other studies have sought to investigate the validity of the tasks through an analysis of score data (Iwashita et al., 2008; Lee, 2006; Sawaki & Nissan, 2009; Stricker & Rock, 2008). In addition to content-based and correlation studies, a further type of evidence pertains to the criteria that are used to judge performance in the two contexts.

While Language for Specific Purposes (LSP) test content and methods are typically derived from an analysis of the target language use situation, the criteria by which performances are judged are seldom derived from the same source but instead derive from 'theoretical understandings of what it means to know and use a language .... without regard, in some cases, for the situation in which it is used' (Douglas, 2001, p. 173). The development of the TOEFL iBT™ scales was based initially on theoretical definitions of communicative competence, and subsequently informed by empirical data consisting of qualitative themes of raters' judgments. However, these judgements were gathered from ESL instructors, oral assessment specialists, applied linguists, and university administrators, and not academic tutors (Cumming et al., 2004; Brown et

al., 2005; Enright, Bridgeman, Eignor, Kantor et al., 2008; Enright, Bridgeman, Eignor, Lee & Powers, 2008; Jamieson & Poonpon, 2013).

There is as yet no research investigating the relationship between criteria used to assess performance on TOEFL iBT™ speaking tasks and those used to assess performance on speaking tasks within academic programs. Such research can contribute to the validity argument by providing evidence that the qualities contributing to scores in the TOEFL iBT™ speaking tasks are indeed valued, and contribute to measures of success, in the academic context. In this study, therefore, we examine academic tutors' assessments of first year undergraduate oral presentations and subsequently draw comparisons between these assessment orientations and those prescribed by the TOEFL criteria.

While we had access to the written rubrics used by the tutors we chose to elicit verbal reports, as verbal reports produced while, or immediately after, rating typically produce a much richer and more nuanced picture of what the raters value in the performance (see, for example, Brown, 2007; Brown et al., 2005; Lumley, 2005; Meiron, 1998). Moreover, given that the tutors had not been trained in the application of the rubrics we can assume that they may interpret them differently (Payne, 2003; Reddy and Andrade, 2010) or, indeed, may resist their use in order to draw on their own views of what constitutes quality. Following our analysis of the verbal report data we map the categories onto those represented in the TOEFL rubrics in order to identify the extent to which the tutors' judgements reflect the same assessment orientation as that of the TOEFL speaking rubrics.

Participants and data

Ten local and ten international students enrolled in a core first-year subject in each of three discipline areas - Business Economics and Law, Health Science, and Science and Engineering - were recruited to participate in the study. Each course contained an oral assessment task where the students were required to prepare and deliver a five to ten-minute presentation on a specified topic or theme. The presentations of the participating students were videotaped for subsequent assessment and use in the verbal report sessions. As the courses were core subjects, with large numbers of students, several tutors were involved in their delivery. For each subject, two of the tutors were recruited to participate in the verbal report sessions.

The three courses were Management information systems (MIS), Social determinants of health (SDH), and Organisation and Function of Cells and Organisms (BIO). For each course there were multiple presentation topics. The MIS tasks required students to research and present a topic area related to the major systems covered in the subject: enterprise resource planning, customer relations management, supply chain management, and information technology security. The SDH topics included the attitudes of health professionals, folk health beliefs, the role of the media, and the

influence of family and friends. The SDH task also required students to end their presentation with a question designed to start a class discussion, and also to manage that discussion. The three BIO topics were *Arabidopsis*; phylogenetics; and *Archaeans*, bacteria and eukaryotes. The rubrics for each of the courses are included as Appendix 1.

The data consist of the verbal reports produced by the two course tutors when discussing the scores they had awarded to students' performances. The reports were collected using immediate retrospective think-aloud methodology, as follows. First the two course tutors watched the video-recording of the presentation and noted down their score, referring to the assessment rubric prescribed for the task. Immediately after this, the tutors participated in the verbal report session, which was structured as follows:

1. The researcher invited each of the tutors to indicate their overall assessment.

2. The researcher invited each of the tutors in turn to justify their score by describing the features they attended to.

The researchers used a script to structure the elicitation of marks and commentary (Figure 1).

Preamble: What we are interested in is the aspects of the performance that you focus on, NOT the score you give. I will ask you for your score simply as a way in to talking about the performance, but if you disagree, which is quite possible, it really doesn't matter - although it might be interesting for you to see what it is that leads to different scores.
A and B. What score did you give for X?
If they agree:
A - Why did you give this score? What did you notice in the presentation?
B - Why did you give this score? What did you notice in the presentation? Did you focus on the same features or any different things?
If they disagree:
A - What did you focus on in the presentation to arrive at your score. Can you justify it to B?
B - You gave a different score. What did you focus on? Did you focus on different things?

Figure 1. Tutor discussion questions

As two raters provided reports in the same session, rather than a single rater, this was not a typical verbal report procedure. While it could be argued that one rater's report could potentially influence the other's, there were two reasons for carrying out the data collection this way. Firstly, as the tutors had not been trained to carry out the oral assessment and were relatively new to it, we felt that paired discussion would encourage them to seek and produce more detail as they sought to justify their scores to each other, without the interviewer becoming involved. Also, by setting the session up as a 'moderation' session, focusing on scores, we also avoided having to be too

explicit about our focus on the performance features and hence possibly constraining what they might refer to.

Analysis of the verbal report data

The verbal reports were transcribed and cleaned of hesitations, repetitions and repairs prior to segmentation and coding. The first step in the analysis was to scan the cleaned data in order to develop an approach to segmentation. We settled on a meaning-based unit of analysis, that is, a single or several contiguous utterances with a single or multiple intertwined aspects of the performance as the focus but which can be separated, both syntactically and semantically, from neighbouring units.

The two researchers began by segmenting six transcriptions (10% of the total) individually. Reliability was calculated, with an average percentage agreement of 83%. Disagreements were discussed until agreement on all segments could be reached. Next, the remaining data were again segmented individually, with all disagreements being subsequently discussed and agreed. The segmented data were entered into one column of an Excel spreadsheet, with identifying data such as student and tutor IDs in additional columns.

The next step of the analysis involved an attempt to specify a set of coding categories using grounded theory, and open coding, using an inductive approach to conceptualise and classify the patterns in the data. After re-reading the entire data-set, we identified tentative categories based on distinctions we could see in the data. This proved to be challenging, for a number of reasons. One general difficulty was that some comments were very broad and, at times, unspecific (e.g., 'That was clear'). At times such comments were clarified through elaborations, but at other times it was difficult to determine the exact reference. As different aspects of the performance were sometimes syntactically interwoven, and hence not separable into discrete segments, it was decided that the most appropriate way to code the data would be to assign codes to each unit for as many performance features as could reliably be identified. The codes were entered into cells alongside the extracts in the Excel spreadsheet. The coding process was iterative: after a single transcript had been coded individually, agreement was calculated, followed by discussion of the differences. This step was repeated six times on different transcriptions (10% of the total), with an average percentage agreement of 82%. The remainder of the data was coded individually by the two researchers and, as before, all disagreements were subsequently discussed in order to reach final agreement.

Some units did not refer to aspects of performance. For example, a number of units were merely statement of the score awarded. These comments were coded as 'other' and subsequently deleted from the analysis.

Findings

Verbal report analysis

We found that eleven independent categories, each concerned with a discrete aspect of performance and together accounting for almost all of the data, could be easily and reliably identified. A twelfth one (Other) consisted of non-performance-related comments or simple statements of scores. Table 1 lists the performance-related categories and the frequency with which they occurred in the data. However, these frequencies should not be interpreted as an indication of their relative importance or salience, given that in their reports the raters necessarily select from, or summarise, the myriad thoughts that consciously pass through their minds as they evaluate the performance in real time, and may be more comfortable with articulating some aspects of performance than others (Lumley, 2005). The categories tend, on the whole, to be quite broad, because of the general nature of many of the comments. So, for example, many comments in the Delivery category referred simply to 'delivery' and not to specific aspects of delivery such as fluency, or pronunciation. In the remainder of this section we present each of the categories, in order of frequency, with, where relevant, a description of the more detailed orientations found within them along with illustrative examples.

Table 1. Rating focus categories by frequency

Category	N	Freq. (%)	Category	N	Freq. (%)	Category	N	Freq. (%)
Content	157	19.2	Affect	56	6.6	Introductions	14	1.7
Delivery	156	19.1	Preparation	42	5.1	Timing	9	1.1
Support materials	136	16.6	Language	33	4.0			
Non-verbal	98	12.0	Discussion	26	3.2			
Engagement	93	11.4						

Content

Comments in this category addressed the topical content of the presentations. While some comments were very general assessments (e.g., 'The information was great'; 'I think she actually knew the topic'), others alluded to more specific aspects of the content, variously encompassing the following: the relevance and comprehensiveness of the content (Extracts 1 & 2), the extent to which students referred to the structure of their presentations (Extract 3), and the organisation and coherence of the content (Extracts 4 & 5). Other comments focused on whether, or how well, various functional moves required or expected by the task, such as definitions, explanations, expression of opinion, and inclusion of examples, were carried out (Extracts 6 & 7).

Extract 1 (BIO, C)[1]: but she didn't really show the relevance of that connected with the essay topic 'Are viruses alive?' That was more what a virus looks like.

Extract 2 (MIS, A): Interestingly enough I got the impression that he wasn't trying, that he hadn't done a lot of work and it was very superficial. He never went into more than a sentence or two. So a bit slack I thought.

Extract 3 (BIO, C): I did actually like that way he set out the structure: 'I'm now going to talk about the introduction, I'm now up to conclusion'. For me that helped with engaging and being able to follow where he was at.

Extract 4 (BIO, C): It wasn't clear what he was talking about. Originally it looked like he was going to be talking about where do we come from, but then he was focusing more on Polynesia and those islands, and so yeah, hence the lower mark there.

Extract 5 (BIO, M): so there was maybe some problems in the organization, the linking of her presentation, that things weren't necessarily flowing very well.

Extract 6 (SDH, S): she … expressed some really great points but not necessarily her opinion.

Extract 7 (SDH, S): and she has incorporated a personal story as well

Another aspect of the content that the tutors commented on related more to academic skills: the extent to which the students incorporated previous discussion, or reading into their presentation. While incorporation of research was valued (Extracts 8 & 9), there was an expectation that students would not simply repeat what they had read (Extract 10).

Extract 8 (BIO, N): She knew what she was talking about and she had obviously done her research.

Extract 9 (SDH, S): I also noted that the good use of the case study. I put here I love how she introduced and used the research example to passionately express her view point so that stood out to me.

Extract 10 (SDH, K): she wasn't reading from plagiarised material. She also incorporated her own thoughts into it.

Delivery

Comments in this category referred to pronunciation, intonation, speed, voice quality, flow, and volume. The tutors also referred more broadly to 'clarity', which seemed to

[1]This comment was made by tutor 'C'.

function as an overarching category (Extract 11). Flow, or fluency, addressed the regularity of delivery (Extract 12). The quality of the delivery was seen as having a huge impact on the extent to which the speaker was able to engage the audience (Extract 13). While students were not specifically prohibited from reading their scripts (and indeed tutors voiced an expectation that at this stage of their course they might resort to reading), the tutors saw this as having a negative impact on the quality of delivery and the engagement of the audience.

Extract 11 (BIO, C): I loved his clarity, his volume, I just found it really easy to listen to.

Extract 12 (BIO, C): She was using a very stop start sort of flow which was very distracting and not engaging at all really.

Extract 13 (SDH,S): I put a judgemental comment in my own notes not that I would ever tell her that but I found that it was boring and I switched off and yet the information was great but it is because there was no variation in her voice.

Support materials

Comments in this category referred to the use of support materials - mainly notes and PowerPoint displays, but also videos, cartoons and handouts. In all the subject areas, students were allowed to use notes when presenting. In two (BIO and MIS) they were also encouraged to use PowerPoint displays or other visual support materials such as video or handouts. A considerable proportion of the tutors' comments were directed at these support materials, with the amount and density of text, the use of interesting visuals, and the layout being evaluated (Extracts 14). Well-designed displays were seen as supporting the audience's understanding (Extract 15).

Extract 1 (BIO, M): Her slides were attractive, they were nicely set out, but I found that on some of them there was a bit too much text, and while they were informative, on some of the slides, also, there was maybe too many concepts and they could have been broken up into more slides.

Extract 2 (MIS, J): And because of the information that he included in the slides, are useful, so you can follow like the whole presentation.

The tutors also attended to the way students used their notes. Over-reliance was evaluated negatively, as resulting in a lack of engagement. It was attributed variously to lack of preparation, lack of understanding of the content, and nervousness (Extracts 16 & 17). Reading from the slides was evaluated similarly.

Extract 3 (SDH, K): She must have practiced but maybe not enough because I felt that in the first part of her presentation she was looking up and as she moved on she got more boring, read more and made much less eye contact.

Extract 4 (BIO, C): and I was kind of thinking, 'Wow, there's your notes over there. When are you going to pick them up, and when are you going to look at them?' But she didn't once. I was very impressed.

Non-verbal behaviour

The comments in this category referred to body language and eye contact. Although the three disciplines differed in the extent to which their rubrics addressed this aspect of performance, tutors in all domains referred to the students' non-verbal behaviours. In particular eye contact was considered important (Extract 18); students who focused on their notes too much were assessed negatively. Body language was typically linked to confidence or nervousness (Extract 19), and sometimes to professionalism (Extract 20). Both eye contact and body language were felt to impact on the engagement of the audience.

Extract 18 (SDH, C): But I felt that she started off well in terms of she must have practiced but maybe not enough because I felt that in the first part of her presentation she was looking up and as she moved on she got more boring, read more and made much less eye contact

Extract 19 (BIO, M): There was a bit of nervous flailing of arms which, as opposed to the girl before who was using her arms as gestures to help sort of keep the audience entertained, I think she was kind of flailing nervously a little bit

Extract 5 (SDH, S): The whole body language I don't know that she was aware she was very casual very relaxed very, I thought, unprofessional, and so I didn't like that.

Engagement

Engagement refers to attracting or holding the audience's interest. Audience engagement was addressed in the rubrics for MIS only, yet it was a major focus of the tutors in all domains. Typically, the extent to which the speaker was able to engage the audience was linked to the use of notes (Extract 21) or the extent to which the PowerPoint display supported understanding - or to the quality of the content (Extract 22) or the delivery (Extract 233).

Extract 6 (SDH, K): If she hadn't looked at her notes so much she would have been more engaging.

Extract 7 (BIO, C): And so it was almost like a bit of, not quite advice giving, but she was using good examples that showed why, and her topic was why health professionals should know about health beliefs and she used personal examples and appropriate examples that were engaging and helped to link those ideas together

Extract 8 (MIS, A): His voice was clear but it was boring. It was a bit of a monotone. He

wasn't engaging the audience.

Affect

Comments in this category contained references to how the student appeared to feel. They are, by their very nature, inferences, typically drawn in relation to the speaker's body language or delivery (Extracts 24 & 25). While the affective state of the speaker was not an assessment focus per se, often the behaviours that gave rise to the inference – such as not making eye contact - were. Where the behaviours were described in a comment along with the inferred affective state, the comments have also been coded in the relevant other category.

Extract 9 (SDH, S): so that could have been nerves and it was just looking down at those notes, not feeling comfortable in looking at that audience and that detracted a little bit

Extract 25 (BIO, C): He was a very natural speaker, very confident, very positive.

Preparation

Comments in this category referred to whether, or the extent to which, the student had prepared for their presentation. These are, of course, inferences that are drawn on the basis of the presentation quality as the tutors could not be expected to know how much planning had gone into individual presentations.

Within the MIS rubrics, preparation is addressed as an aspect of delivery, affecting fluency whereas in SDH it is seen as impacting on the students' ability to meet the time constraints. While such orientations were evident in some comments, (Extract 26), far more often tutors commented on preparation in relation to the quality of the content (Extract 27).

Extract 10 (SDH, S): but she did stumble a fair bit so for me that linked in with also perhaps a lack of planning and rehearsal so that is over those two points I don't think that the standard was met where it should be.

Extract 11 (SDH, K): She obviously was well prepared, she'd done a little bit of research.

Language

Comments in this category tended to address vocabulary or the level of formality of the language used in the presentation. In terms of vocabulary, the tutors focused on the general vocabulary range of the speaker (Extract 28) or the use of technical or subject-specific vocabulary (Extract 29). Other comments reflected a concern with the register of the speech, that is, whether the student used a spoken or written style, which was typically linked to some students' tendency to read a written presentation

aloud (Extract 30).

Extract 28 (BIO, M): Very, very capable student, very bright girl, and I think that was pretty evident that she speaks really well, and she's got good vocabulary and all that sort of thing.

Extract 29 (BIO, M): But overall she was actually first semester student and it's just my opinion that that topic is actually quite hard to talk about, it's quite – you know I find it quite hard to talk about that and she was struggling over a lot of the words.

Extract 12 (SDH, K): But I felt the language that she used was probably more appropriate to the year level so she had clearly done her reading but she spoke using more sorts of common language than maybe the first student had which was read out in a way that she might have written an essay for example so she used language that you might use for an essay.

There were a handful of comments on the non-nativeness of some students' language, indicating an awareness of their 'foreignness' but typically seeming to make allowances for this (Extract 31 & 32).

Extract 13 (MIS, A): … although the fact that she was an overseas student I was a little bit more generous with her because her language was still quite good. It had no ups and downs but I could still understand it and that's a good start, so I'm always a bit more generous to the overseas students.

Extract 14 (MIS, J): I was also concerned because he was reading too much but he was, I think, I guess trying to help himself with the language barriers so it is the reason why he was talking slowly.

It is notable that despite language being an assessment focus in all three rubrics, there were remarkably few comments in this category.

Discussion

Comments in this category were limited to SDH presentations only and were task-related in that the SDH students were required to end their presentation with a question designed to start a class discussion, and to manage that discussion. Not unsurprisingly, the SDH tutors commented on whether, or how well the students did this, focusing on how interactive is was and how well they responded to other students. A common criticism was that they simply asked a question and did not go further in generating discussion.

Extract 33 (SDH, S): I think that she asked a question but I felt like it was for the sake of asking a question and then she didn't convincingly show any interest in what the student was saying and that was such a pity because at that point that is where the

task is about generating this discussion.

Introductions

Comments in this category made reference to whether, or how well, the students introduced themselves or the group, and the topic (Extract 34). While the SDH rubric was the only one that referred to introductions, we found comments in this category from both SDH and MIS tutors. The comments indicated that the inclusion of an introduction was a necessary listener-oriented requirement.

Extract 34 (MIS, A): First of all I thought her topic was security, and then I realised it was CRM and I felt she didn't link it to start with so that actually took my point five off because she should have really told us where she was going at the beginning so we knew it was CRM the aspect, what they were going to look at.

Timing

Comments in this category referred to the extent to which the speaker had fulfilled the task requirement pertaining to timing. Adherence to the time limit imposed on the task appeared as an assessment focus in the SDH and MIS rubrics, and was addressed by tutors from these subjects only.

Comparison with TOEFL rubrics

In the analysis of the verbal reports we found that the tutors oriented to all three aspects of performance addressed in the TOEFL rubrics (Educational Testing Service, 2019) – content, delivery and language - but with some major differences. Firstly, the academic assessments showed a stronger commitment to the presentation as a measure of *academi*c skills. Content was judged not only in terms of the ability to produce relevant, well-organized and meaningful content, as is the case in the TOEFL rubrics, but also the extent to which it demonstrated that the student had done the research and understood the topic, which is something that is clearly not possible in TOEFL given that tasks are removed from the academic context. This perspective concurs with that of Joughin (1998: 368), who describes academic oral assessment as being where 'the object of assessment is not the oral ability of the student but rather the student's cognitive knowledge, understanding, thinking processes, and capacity to communicate in relation to these'.

Secondly, the academic assessments showed a stronger commitment to presentation skills, arguably another academic skill as a typical graduate competency, than is the case in TOEFL. Audience awareness was invoked or implied as being valued in many of the comments across all categories, for example, the selection of content that is engaging or relevant to the listeners' needs, eye contact with the audience, the students' professionalism or demeanour while presenting (cf. Douglas & Myers, 2000),

and the use of visual support materials to support understanding (as has been found in analyses of academic presentations in other studies, e.g. Jacoby & McNamara, 1999; Raof, 2011). All of these can be seen as aspects of the student's socialisation into academic culture (cf. Duff, 2007, 2010).

Language proved to be the focus of attention surprisingly little of the time, and while the majority of the comments on language were directed at non-native-speakers, for many of these students there were no such comments at all. When they did attend specifically to language, the tutors tended to comment on the students' use of appropriate technical vocabulary, on the one hand, and the orality of their speech – the register – on the other. Other comments were not interpretable as they addressed the quality of 'spoken language' more generally, and may indicate a lack of linguistic awareness on the part of the tutors. In fact, the rubrics used by the tutors had very little detail on language per se, and thus did not provide them with the means to take a more nuanced approach to the assessment of language. For example, unlike the TOEFL rubrics there was no reference to grammar and syntax or to range or accuracy of structures. Not surprisingly, there was no reference to these by the tutors.

Of course, we must also bear in mind that the adequacy of the language may not be an issue in these contexts, for the native-speakers and possibly also for the non-native-speakers, who have, after all, already been admitted to the university. While this is a limitation of our study, it is in our view by no means the case that all the local students, let alone the international students, were fluent, or even competent and confident speakers, and where there were limitations in some students' language these did not receive great attention and nor were the problems analysed by the tutors in any depth. This contrasts markedly with the rich analysis of student language in verbal reports produced by university-based *language* specialists when evaluating TOEFL speaking performances in the study by Brown et al. (2005).

Conclusion

Analyses of assessment activities indigenous to a particular context can throw light on the often inexplicit criteria which drive insiders' perceptions of communicative performance. This study provides an exploration of the criteria by which oral presentation tasks contributing to student assessment within academic programs are judged. In doing so, it seeks to add to the body of evidence addressing the TOEFL iBT™ extrapolation inference of its validity argument.

On the whole, our findings were mixed. We found that there were some similarities in focus within the two contexts, across the content and delivery of the speech, but that the *quality* of the language appeared to be much less of a focus in the academic context even where, in our view as language specialists, students were struggling to express themselves. Not only were there very few comments on the language, whereas the

TOEFL rubrics weight language equally with delivery and content, but consideration of language appeared to be restricted to vocabulary and register, with seemingly no consideration of grammar and syntax. Language – and a narrow view at that - is only one aspect of oral communication success in the academic context, and from our study does not appear to be the most salient aspect, which supports the argument by Bridgeman et al. (2016, p. 308) that 'English language skills are a necessary but not sufficient condition for success in academic study for international students'. It also, quite possibly, reflects the fact that the subject specialist tutors in our study were untrained in the assessment of the oral presentations, and unguided in how to assess language.

We also found a much greater focus on the performance as a communicative event. Perhaps what our study tells us is more about what other communication skills, beyond mere control of the language, first-year non-native-speaker students will be expected to demonstrate. Unsurprisingly, also, there was a greater focus on independent academic research and presentation skills – components of academic learning and graduate capabilities – which students are expected to acquire in first year core courses that mark the transition from secondary to tertiary education, a finding that concurs with other studies investigating academic presentations at other, more advanced, levels of academia levels (e.g. Bhati, 2012; Jacoby & McNamara, 1999, Kerby & Romaine, 2009; Moni et al., 2005; Pierce & Robisco, 2010).

Given the differences in the two contexts, not least the embeddedness of the academic presentations in the broader study context and the physical proximity of the players - speaker, audience, and assessor - some criteria used in the academic context will necessarily be irrelevant or unfeasible in the TOEFL context. Content knowledge, which plays an important role in the assessments in the academic context, cannot be included in assessments such as TOEFL iBT beyond that which is given to all students as input, as it would compromise fairness in the measurement of students' abilities. Moreover, even if it were possible to embed a broader range of academic and presentation skills into the TOEFL tasks, this would require a confrontation with what Jacoby and McNamara (1999, p. 235) describe as 'the thorny issue of recontextualizing assessment criteria across boundaries of professional expertise'. It is neither for language specialists to derive more general assessment criteria from specific context- and content-based comments made in an indigenous assessment setting, nor of course, is it a simple matter for language specialists, as assessors, to judge performances in the same way as academic specialists.

So while the content of both assessment procedures is apparently similar in that they involve the presentation of academic material, at least as pertains to the integrated tasks in the TOEFL iBT™ speaking test, communication is understood more broadly in the academic context, in contrast to the way it is conceived within the TOEFL. In other words, the skills evaluated are quite different, arising from the purpose of the

task and the context of the presentation, and the two contexts differ consequently in the criteria used to evaluate the performance. We note, however, that while it was not possible in our study to disentangle the effect of task design on the evaluation criteria, the study was nevertheless able to address the extrapolation inference, with the academic context serving as an appropriate criterion.

While the inclusion of TOEFL rater data might arguably have strengthened the comparison of the two assessment contexts, this was not available to us. However, we could argue that for a test such as TOEFL, raters are well-trained to adhere to the criteria and we can therefore assume that the written rubrics are closely adhered to. Ratings of oral academic presentations are far newer and less well known, hence our primary focus on data derived in this context. Finally, we note the value of undertaking what are essentially very complex and labour-intensive analyses in the validation of high-stakes tests such as TOEFL iBT™. Such data provide helpful insights into the indigenous academic practices, and serve to complement other types of evidence such as score or discourse data. They can also, as we have seen here, highlight the limits of convergence between the two contexts.

Acknowledgement: This research was funded by the Educational Testing Service (ETS) under a Committee of Examiners and the Test of English as a Foreign Language research grant. ETS does not discount or endorse the methodology, results, implications, or opinions presented by the researcher(s).

References

Barrie, S. Hughes, C. Crisp, G., & Bennison, A. (2012). Assessing and assuring graduate learning outcomes: project summaries. The University of Sydney, The University of Queensland and RMIT University. Available online: http://sydney.edu.au/education-portfolio/qa/projects/aaglo/pdf/SP10-1879_FINAL%20sydney%20barrie%20final%20report%20part%201.pdf

Bearman, M., Dawson, P., Boud, D., Bennett, S., Hall, M., & Molloy, E. (2016). Support for assessment practice: developing the Assessment Design Decisions Framework. *Teaching in Higher Education, 21*(5), 545-556. DOI: 10.1080/13562517.2016.1160217

Bhati, S. (2012). The effectiveness of oral presentation assessment in a finance subject: an empirical examination. *Journal of University Teaching & Learning Practice, 9*(2), 1-22. https://ro.uow.edu.au/jutlp/vol9/iss2/6

Biber, D. (2006). *University language: a corpus-based study of spoken and written registers.* Amsterdam: John Benjamins.

Biber, D., Conrad, S. M., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E., & Urzua, A. (2004). *Representing language use in the university: analysis of the TOEFL® 2000 spoken and written academic language corpus.* (TOEFL Report

Monograph Series 25). Princeton, NJ: Educational Testing Service. https://www.ets.org/research/policy_research_reports/publications/report/2004/ibyq

Bowden, K. (2010). Background paper for the AQF Council on generic skills. South Australian Department of Further Education Employment Science and Technology. Downloaded from http://hdl.voced.edu.au/10707/166337

Bridgeman, B., Cho, Y., & Di Pietro, S. (2016). Predicting grades from an English language assessment: The importance of peeling the onion. *Language Testing*, *33*(3), 307-318. https://doi.org/10.1177/0265532215583066

Brooks, L. & Swain, M. (2014). Contextualizing performances: comparing performances during TOEFL iBT™ and real-life academic speaking activities. *Language Assessment Quarterly, 11*(4), 353-373. DOI: 10.1080/15434303.2014.947532

Brown, A. (2007). An investigation of the rating process in the IELTS oral interview. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 98-139). Cambridge: Cambridge University Press.

Brown, A. & Ducasse, A. (2019). An Equal Challenge? Comparing TOEFL iBT™ Speaking Tasks with Academic Speaking Tasks. *Language Assessment Quarterly, 16*(2), 253-270. DOI: 10.1080/15434303.2019.1628240

Brown, A., Iwashita, N., & McNamara, T. F. (2005). *An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks.* (ETS Research Report Series 29, RR-05-05). Princeton, NJ: Educational Testing Service. https://www.ets.org/research/policy_research_reports/publications/report/2005/hsiw

Chapelle, C., Enright, M., & Jamieson, J. eds. (2008). *Building a validity argument for TOEFL*. New York, NY: Routledge.

Cumming, A., Grant, L., Mulcahy-Ernt, P. & Powers, D. (2004). A teacher-verification study of speaking and writing prototype tasks for a New TOEFL. *Language Testing, 21*(2), 107-145. https://doi.org/10.1191/0265532204lt278oa

Douglas, D. (2001). Language for Specific Purposes assessment criteria: Where do they come from? *Language Testing, 18*(2), 171-185. https://doi.org/10.1177/026553220101800204

Douglas, D., & Myers, R. (2000). Assessing the communication skills of veterinary students: Whose criteria? In *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida*, edited by A. J. Kunnan , 60-81. Cambridge: University of Cambridge.

Duff, P. A. (2007). Problematising academic discourse socialisation. In *Learning discourses and the discourses of learning, 1, 1*, edited by Marriott, H., Moore, T., & Spence-Brown, R., 1-18. Melbourne & Sydney: Monash University ePress / University of Sydney Press.

Duff, P. A. (2010). Language socialization into academic discourse communities. *Annual review of applied linguistics*, *30*, 169-192. DOI: 10.1017/S0267190510000048

Educational Testing Service. (2019). TOEFL iBT Test Speaking Rubrics. Princeton, NJ: Educational Testing Service. https://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf

Enright, M. K., Bridgeman, B., Eignor, D., Kantor, R. N., Mollaun, P., Nissan, S., Powers, D. E., & Schedl, M. (2008). Prototyping new assessment tasks. In *Building a Validity Argument for the Test of English as a Foreign Language,* edited by Chapelle, C. A., Enright, M. K. and Jamieson, J. M., 97-143. New York and Oxford: Routledge.

Enright, M. K., Bridgeman, B., Eignor, D., Lee, Y.-W., & Powers, D. E. (2008). Prototyping measures of Listening, reading, speaking and writing. In *Building a Validity Argument for the Test of English as a Foreign Language,* edited by Chapelle, C. A., Enright, M. K., & Jamieson, J. M., 145-186. New York and Oxford: Routledge.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*, 24-49. https://doi.org/10.1093/applin/amm017

Jamieson, J., & Poonpon, K. (2013). *Developing analytic rating guides for TOEFL iBT® integrated speaking tasks.* (ETS Research Report Series 13-13, TOEFLiBT-20). Princeton, NJ: Educational Testing Service. https://www.ets.org/research/policy_research_reports/publications/report/2013/jqoc

Jacoby, S., & McNamara, T. F. (1999). Locating competence. *English for Specific Purposes*, *18*(3), 213-241. https://doi.org/10.1016/S0889-4906(97)00053-7

Joughin, D. (1998). Dimensions of oral assessment. *Assessment & Evaluation in Higher Educatio*n, *23*(4), 367-378. https://doi.org/10.1080/0260293980230404

Kerby, D., & Romaine, J. (2009). Developing oral presentation skills through accounting curriculum design and Course-Embedded Assessment. *Journal of Education for Business, 85*(3), 172-179. https://doi.org/10.1080/08832320903252389

Lee, Y. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing*, *23*(2), 131-166. DOI: 10.1191/0265532206lt325oa

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt am Main: Peter Lang.

Meiron, B. E. (1998). Rating oral proficiency tests: a triangulated study of rater thought processes. Unpublished master's thesis, University of California at Los Angeles.

Moni, R. W., Beswick, E., & Moni, K. B. (2005). Using student feedback to construct an assessment rubric for a concept map in physiology. *Advances in Physiology Education*, *29*(4), 197-203. https://doi.org/10.1152/advan.00066.2004

O'Sullivan, A. J., Harris, P., Hughes, C. S., Toohey, S. M., Balasooriya, C., Velan, G., & McNeil, H. P. (2012). Linking assessment to undergraduate student capabilities

through portfolio examination. *Assessment & Evaluation in Higher Education*, *37*(3), 379-391, DOI: 10.1080/02602938.2010.534766

Payne, D. A. (2003). *Applied educational assessment.* Belmont, CA: Wadsworth/ Thomson Learning.

Pereira, D., Flores, M. A., & Niklasson, L. (2016). Assessment revisited: A review of research in assessment and evaluation in higher education. *Assessment & Evaluation in Higher Education*, *41*(7), 1008-1032. DOI: 10.1080/02602938.2015.1055233

Pierce, J. & Robisco, M. (2010). Evaluation of oral production learning outcomes for higher education in Spain. Assessment & Evaluation in Higher Education, *35*(6), 745–758. DOI: 10.1080/02602930902977764

Raof, A. H. A. (2011). An alternative approach to rating scale development. In Language testing: theories and practices, edited by O'Sullivan, B., 151-163. Hampshire: Palgrave Macmillan.

Reddy, Y.M., & Andrade H. (2010) A review of rubric use in higher education. *Assessment & Evaluation in Higher Education, 35*(4), 435-448. DOI: 10.1080/02602930902862859

Rosenfeld, M., Leung, S., & Oltman, P. K. (2001). *The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels*. (TOEFL Monograph Series 21, RM-01-03,). Princeton, NJ: Educational Testing Service.

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. Assessment & Evaluation in Higher Education, *34*(2), 1-22. DOI: 10.1080/02602930801956059

Sawaki, Y., & Nissan, S. (2009). *Criterion-related validity of the TOEFL® iBT listening section.* (ETS Research Report 09-02, TOEFLiBT-08). Princeton, NJ: Educational Testing Service.

Stricker, L. J., & Rock, D. A. (2008). *Factor structure of the TOEFL internet-based test across subgroups.* (ETS Research Report 08-09, TOEFLiBT-04). Princeton, NJ: Educational Testing

Yi, J. (2012). Comparing strategic processes in the iBT speaking test and in the academic classroom. Thesis Submitted for the Degree of Doctor of Education in Applied Linguistics and TESOL, University of Leicester.

Appendix 1 Discipline Criteria

Biology Oral Presentation feedback

Standard exceeded	*Standard met*	*Standard not met*
Spoken language		
Can be heard by audience	Can be heard by audience	Cannot be heard by audience
Appropriate words/terminology used	May be some use of inappropriate words/terminology (e.g. slang)	Does not use appropriate words/terminology (e.g. slang, non-scientific)
Well delivered & enhances understanding	Delivery does not impede understanding	Delivery impedes understanding
Body language		
Faces audience	Mostly faces audience	Does not face audience
Eye contact with audience most of the time	Attempts to make eye contact with audience	No eye contact with audience
Minimal reliance on prepared text	Some reliance on prepared text	Relies completely on prepared text
Purpose & Structure		
Communicates key message clearly – logical order & appropriate level for audience	Mostly communicates key message - unclear in parts & may not always be at appropriate level for audience	Does not communicate key message clearly – no logical structure & may not be at appropriate level for audience
Supporting material mostly integrates with presentation	Supporting material does not distract from presentation	Supporting material does not integrate with presentation

Management Information Systems Marking Guide

Content	*Quality of the research you've done*
Argument	The way in which you integrate the material into a coherent, well-argued exposition of your topic.
Effective use of evidence	The way you've used your research to support your argument.
Quality of Presentation Skills (Criteria below)	How effectively you communicate.

Presentation Skills	*Standard NOT MET*	*Standard MET*	*Standard EXCEEDED*

Structure	Incoherent structure Few relevant points	Identifiable structure, but may not be consistent	Logical, coherent structure throughout
Use of technical, academic and/or professional language	Insufficient or inappropriate use of simple technical or academic vocabulary.	Technical language used correctly but may be somewhat limited. More complex technical language used with few errors. Academic language is used appropriately throughout	Relevant technical language is used correctly throughout. New or complex technical terms are explained effectively to classmates. Academic language used appropriately and integrated throughout.
Audience engagement	Little attempt to make content/presentation style relevant to audience. Classmates unable to identify purpose of the presentation.	Generally, attempts to make presentation relevant to audience. Classmates generally able to identify the key points in the presentation.	Presentation consistently targets the needs and/or interests of classmates. Classmates consistently engage with all aspects of the presentation.
Presentation delivery	Speaker hesitant or unprepared. (In group presentations) inadequate contribution by some presenters or one presenter dominates. Lack of verbal fluency disrupts classmate or teacher understanding.	Generally, well prepared and straightforward presentation with few hesitations, however may lack originality. (In group presentations) evidence that all participants equally involved. Verbal fluency adequate for classmate understanding.	Confident delivery with clear evidence of preparation and awareness of audience needs. (In group presentations) all participants equally participate, in a well-integrated way. Verbal fluency aids classmate understanding throughout presentation.

Social Determinants of Health Speaking Skill Assessment Table

		Standard not met	*Standard met*	*Standard exceeded*
The Speaker....	Introduces self and topic.		Yes ☐ No ☐	
1.	Speaks clearly in English - speech is well paced and audible.	☐	☐	☐
2.	Focuses on the purpose of the task.	☐	☐	☐
3.	Successfully transmits knowledge, ideas and information.	☐	☐	☐
4.	Demonstrates understanding and interest in the material presented.	☐	☐	☐
5.	Uses language and vocabulary that has meaning to both self and audience.	☐	☐	☐
6.	Responds effectively to, and/or uses, non-verbal cues, prompts or visual tools effectively, and the audience.	☐	☐	☐
7.	Meets time constraints showing evidence of planning and rehearsal	☐	☐	☐
Overall evaluation		☐ 1 or more ticks are at this level.	☐7 ticks appear at this level or above	☐ 4 or more ticks are at this level & remainder are at 'standard met' level.