Article

LANGUAGE TESTING

Language Testing
1–20

DOI: 10.1177/0265532214544393
ltj.sagepub.com
SAGE

Coming to terms with innovative high-stakes assessment practice: Teachers' viewpoints on assessment reform

Martin East
University of Auckland, New Zealand

Abstract

Implementing assessment reform can be challenging. Proposed new assessments must be seen by stakeholders to be fit for purpose, and sometimes the perceptions of key stakeholders, such as teachers and students, may differ from the assessment developers. This article considers the recent introduction of a new high-stakes assessment of spoken proficiency for students of foreign languages in New Zealand high schools. The new assessment aims to measure spoken proficiency through the recording of a range of unstaged peer-to-peer interactions as they take place throughout the year. It contrasts with an earlier assessment that drew on a summative teacher-led interview. The article presents findings from a survey of teachers (n = 152), completed two years into the assessment reform, in which teachers were asked to consider the relative usefulness of the two assessment formats. Findings suggest that teachers consider the new assessment to be, in most respects, significantly more useful than the earlier model, and that the new assessment is working relatively well. Some challenges emerge, however, in particular around the feasibility and fairness of collecting ongoing evidence of spontaneous peer-to-peer performance. Findings raise issues to be considered if the new assessment is to work more successfully.

Keywords

Interactional competence, interview tests, paired speaking tests, spoken proficiency, test usefulness

Implementing assessment reform can be a tricky business. Challenges to implementation lie in the reality that the real worlds of stakeholders may be very different to the theoretical world in which assessment developers anticipate that their proposed assessment will

Corresponding author:
Martin East, School of Curriculum and Pedagogy, Faculty of Education, The University of Auckland, Private Bag 92601, Symonds Street, Auckland 1150, New Zealand.
Email: m.east@auckland.ac.nz

be used. As Bachman and Palmer (2010) argue, "people generally use language assessments in order to bring about some beneficial outcomes or consequences for stakeholders as a result of using the assessment and making decisions based on the assessment" (p. 86). Assessment developers therefore generally propose and design assessments with this aim in mind. However, assessments planned with the best of intentions may not always work out as intended, leading to "the possibility that using the assessment will not lead to the intended consequences, or that the assessment use will lead to unintended consequences that may be detrimental to stakeholders" (p. 87).

Seen in the light of the now well-established qualities of test usefulness articulated by Bachman and Palmer (1996), proposed new assessments must be seen by stakeholders to be "useful" or "fit for purpose". When the perceptions of the primary consumers (teachers and students) differ from the assessment developers, very strong feelings about the assessment can be evoked.

This article presents one case of assessment reform – the recent introduction of a new high-stakes assessment of spoken proficiency for students of foreign languages (FLs) in high schools in New Zealand. The new assessment, known as *interact*, is intended to capture for assessment purposes a series of genuine and unrehearsed student-initiated peer-to-peer spoken interactions. It stands in contrast to an earlier assessment, *converse*, which relied on a summative one-time conversation between teacher and student. The primary purpose of *interact* is therefore to move teachers and students away from assessments of spoken proficiency that had effectively become one-sided engagements in somewhat staged "conversations", and towards assessments that will demonstrate a level of ability to interact spontaneously, negotiate meaning and demonstrate a genuine ability to communicate.

As East and Scott (2011a, 2011b) explain, the implementation of *interact* has not been unproblematic. Teacher feedback from initial consultations, collected during the development stage and before the assessment had been trialled and introduced, suggested considerable teacher anxiety about the negative consequences of introducing the new assessment – for example, significantly increasing teachers' (and students') workloads, potentially disincentivizing some students and aggravating student attrition from year to year.

With a view to establishing, two years into its introduction, how teachers perceive the assessment as it begins to be employed in the real worlds of local classrooms, this article presents findings from data gathered from a national survey. Teachers were asked to compare *interact* with *converse.* Findings are used to determine the relative perceived usefulness (and therefore perceived validity or fitness for purpose) of the two assessments.

Although, as Winke (2011) acknowledges, teacher perspectives are not normally invited as part of considerations about the validity of an assessment, in the case of *interact* the teachers are the principal decision-makers with regard to what will be assessed and how it will be assessed (i.e., the interactions students will carry out for assessment purposes), including, if teachers wish, opting out of the assessment altogether. Furthermore, the choices teachers make about what they do in their classrooms are "underpinned by beliefs about the nature of language, the nature of the learning process and the nature of the teaching act" (Nunan, 2004, p. 6). Borg (2006) suggests that

teachers do not operate as "mechanical implementers of external prescriptions" (p. 7). Rather, teachers are "active, thinking decision-makers who play a central role in shaping classroom events" (p. 1). Asking teachers what they think about the assessments provides a window into their own beliefs and understandings about effective language pedagogy and, following from that, what, in their eyes, constitutes effective assessment.

With regard to *interact* in comparison with *converse*, teachers are uniquely placed to provide valuable information about relative usefulness and are thereby able to "shed light on the validity of the tests, that is, whether the tests measure what they are supposed to and are justified in terms of their outcomes, uses, and consequences" (Winke, 2011, p. 633).

Speaking assessments

The single candidate interview represents the most typical way of operationalizing speaking assessments (Luoma, 2004). In the FL context, arguably one of the most influential examples is the Oral Proficiency Interview test of the American Council on the Teaching of Foreign Languages (ACTFL-OPI). From a theoretical perspective, the ACTFL-OPI is designed to draw attention to candidates' functional use of language in a way that supports "communicative language teaching, with its emphasis on meaningful interaction in the language as opposed to knowledge of linguistic rules" (Turner, 1998, p. 192). The test has had significant washback into FL classrooms, at least in the United States (Yoffe, 1997).

The validity of the ACTFL-OPI has been called into question in a number of ways. Although ostensibly aiming to capture interaction between the test candidate and the examiner, effectively the test is one-sided: the examiner is in control of what happens in the interaction and the candidate responds to the questions the examiner chooses to pose. Also, although, according to Yoffe (1997), the interview "purports to assess functional speaking ability" there is strong encouragement for raters to "pay careful attention to the form of the language produced rather than to the message conveyed" (p. 5), consequently leading to examiner attempts to "force" particular grammatical structures into use. As Kramsch (1986) argued almost forty years ago, this kind of testing stresses "behavioural functions and the lexical and grammatical forms of the language" and overlooks the "dynamic process of communication" (p. 368).

A weakness of single interview tests is therefore that they do not represent normal conversation (van Lier, 1989). If we wish to measure spoken interaction in a valid way, we need to include "reciprocity conditions" (Weir, 2005), or, to use Kramsch's (1986) terminology, the opportunity to measure "interactional competence". The use of paired or group oral interactions as an alternative to teacher- or examiner-dominated interview tests has been growing since the 1980s (Ducasse & Brown, 2009). Paired speaking tests are now commonly used in both high-stakes and classroom assessment contexts (May, 2011).

Peer-to-peer speaking assessments have been found to be more balanced between the interlocutors (Együd & Glover, 2001), and to elicit a wider spectrum of functional and interactional competences such as collaboration, cooperation and coordination (Jacoby & Ochs, 1995), prompting, elaboration, finishing sentences, referring to a partner's ideas

and paraphrasing (Brooks, 2009), turn taking, initiating topics and engaging in extended discourse with a partner rather than an examiner (Ducasse & Brown, 2009; May, 2011). Paired speaking assessments therefore reflect a more comprehensive communicative spoken proficiency construct. This allows for better inferences to be made about the candidate's proficiency in wider real-life contexts, contributing to the validity of the assessment (Galaczi, 2010).

One important consequential advantage of the paired format is the claim of positive washback on classroom practices (Ducasse & Brown, 2009; Messick, 1996). Paired assessments reflect a strong relationship between teaching and assessment by encouraging more pair work in class or reflecting what is already happening in communicatively oriented classrooms (Galaczi, 2010; Taylor, 2001). It has also been asserted that students view paired or group interactions positively and that they are less anxiety-provoking for candidates (Együd & Glover, 2001; Fulcher, 1996; Nakatsuhara, 2009; Ockey, 2001). Paired or group assessments may also be more time and cost efficient because candidates are assessed together, and raters assess two or more candidates simultaneously (Ducasse & Brown, 2009; Galaczi, 2010).

A major limitation of the paired format is whether it matters who is paired with whom (Foot, 1999; Galaczi & ffrench, 2011). So-called "interlocutor effects" (O'Sullivan, 2002) such as age, gender, cultural or first language background, personality, status or degree of familiarity or acquaintanceship can influence the amount and quality of the discourse. Although studies of the impact of pairing candidates have led to contrasting findings (Csépes, 2002; Nakatsuhara, 2011; Norton, 2005), interlocutor variables can potentially become threats to the test's validity and fairness (Galaczi & ffrench, 2011).

An allied concern relates to scoring. Participants' performances are co-constructed and interdependent (Brooks, 2009), whereas we often require measures of individual performance (McNamara & Roever, 2006). If, as Weir (2005) contends, "an individual's performance is clearly affected by the way the discourse is co-constructed by the person they are interacting with", this becomes a problem for the reliable measurement of individual proficiency, and yet "[h]ow to factor this into or out of assessment criteria is yet to be established in a satisfactory manner" (p. 153). There are several genuine and potentially negative consequences for candidates taking part in an asymmetric interaction (May, 2009).

When considered with regard to the relative usefulness of the established single-candidate interview in comparison with paired (or group) speaking assessments, it may be concluded that the jury is still out. With this in mind, this article addresses teachers' perceptions of relative usefulness in a context where, as a consequence of curriculum and assessment reform, the former assessment type is being replaced with the latter.

Background

The start of the 21st century has witnessed considerable curriculum and assessment reform in New Zealand. A new curriculum for schools (Ministry of Education, 2007) was fully implemented from the start of the academic year 2010. Building on a growing encouragement for schools to embrace a sociocultural pedagogy, the revised curriculum

is intended to facilitate a learner-centred and experiential approach in contrast to a top-down, teacher-led model.

The introduction of a revised school curriculum gave rise to a subject-wide review of New Zealand's national high-stakes assessment system for secondary schools, the National Certificate of Educational Achievement (NCEA). The review was to ensure that the individual components of NCEA, known as "standards", were aligned with curricular aims and intentions. Between 2008 and 2010 the original NCEA standards, the blueprints from which specific assessment tasks are developed (Bachman & Palmer, 2010) and which represent the "skills or knowledge that [students] are expected to achieve or know" in a selected area of learning (NZQA, 2009), were rewritten, consulted on, and then revised in light of the feedback.

Revised NCEA standards have been progressively introduced: level 1, taken by students at the end of Year 11 (age 15+ and final year of compulsory schooling) was implemented in 2011. Level 2 (Year 12) was phased in in 2012, and level 3 (Year 13, final year of voluntary schooling) in 2013. Thus, teachers have recently had to come to terms with both a substantially revised whole school curriculum and aligned high-stakes assessments in a rolling programme of implementation.

With regard to FL programmes, curricular innovation has provided strong encouragement for teachers to move away from teacher-led systematic and structured notional-functional or oral-situational syllabi that have become well established in New Zealand's schools, and to consider the contributions of more learner-centred, open-ended approaches such as task-based language teaching (Rahimpoura, 2010). As a consequence, a strong emphasis of New Zealand's revised FL curriculum is on students "learning to communicate through interaction in the target language" (Nunan, 2004, p. 1). The proposal to replace *converse* with *interact* reflected revised curricular aims.

The *converse* test, influenced by a traditional behaviourist product-oriented and knowledge-based approach to assessment (East, 2008), required a single static summative teacher-conducted interview. In line with the more traditional (teacher-led and grammar-based) communicative approach to teaching and learning then in play, evidence was required, and therefore sought, that candidates could use specified grammatical structures commensurate with their stage in learning (NCEA levels 1, 2 or 3). In practice, students often drew on a good deal of scripted and pre-learnt material (even though the assessments were supposed to include an element of "unpredictable" language). Evidence of interaction was therefore often incomplete, contrived, inauthentic and accuracy-dominated.

By contrast, *interact* aims to reflect a constructivist process-oriented dynamic assessment model (East, 2008) and was designed, in theory at least, to capture a range of genuine and unrehearsed student-initiated opportunities for peer-to-peer interactions as they took place in the context of ongoing classroom work. This was not meant to preclude teachers creating dedicated assessment opportunities. It was, rather, meant to facilitate the ongoing collection of authentic and direct samples of speaking (Messick, 1996).

Teachers would be required to use three instances of interaction for summative grading purposes. It was anticipated that these would be selected from recordings of a range of interactions that took place throughout the year. Teachers would assess the three samples holistically and decide on the appropriate level of performance based on

the assessment criteria and the overall evidence presented (see Appendix A). To ensure consistent application of the standard across schools, samples of work would be moderated internally by colleagues and made available on request to an external moderator. Informed by a more open-ended programme of learning, the active use of specified grammatical structures was no longer required in the assessment, although it was anticipated that accuracy would be indirectly evaluated in terms of its contribution to communicative effectiveness.

A range of resources was made available to support teachers with the introduction of *interact* and application of the grading criteria. These included the following: examples of tasks (TKI, 2014a, 2014b); workshops where teachers could work with real samples of students' work and engage in discussion about interpreting the standard, thereby increasing their confidence about making appropriate assessment judgments (NZQA, 2014a); annotated audio examplars of student performances at a range of levels (NZQA, 2014b); and periodical clarification documents that would address issues as they arose and would help to clarify the published expectations of the assessment (NZQA, 2014c).

The present study

From the perspective of its designers, it was anticipated that *interact* would promote more valid assessment opportunities than *converse* in line with curricular expectations. However, as previously stated, initial teacher feedback during the development stage suggested substantial concerns (East & Scott, 2011a, 2011b). The present study, wide-ranging in scope and drawing on a range of data sources, was implemented to investigate stakeholders' perspectives on the assessment reform during the period of its roll-out (2012–2013). This article reports on one aspect of Phase I of the study, which took place towards the end of 2012 (two years into the reform and after NCEA levels 1 and 2 had come on stream). It draws on data from a large-scale nationwide anonymous paper-based survey, targeted at teachers of the five principal FLs taught in New Zealand (Chinese, French, German, Japanese and Spanish). The survey sought teachers' perceptions of *interact*, whether or not they had chosen to use the new assessment since its introduction.

The following primary research question is addressed: To what extent are *converse* and *interact* supported by language teacher perceptions? The sub-question addressed is as follows: To what extent are *converse* and *interact* perceived as useful?

Design

Section I of the survey consisted of 10 paired statements, one referring to *converse* and the other to *interact*. Section II consisted of four open-ended questions. (Survey questions are reproduced in Appendix B.) The overarching construct measured by the statements (all 10 paired items) was perceived usefulness. This construct was interpreted as incorporating six qualities (Bachman & Palmer, 1996): construct validity (whether the assessment adequately reflects the construct of spoken proficiency that the assessment is intended to measure); reliability (whether performances can be consistently and accurately measured); authenticity (whether the assessment adequately reflects spoken target

language use in real world domains beyond the assessment); interactiveness (whether students can engage meaningfully with the assessment task); impact (whether the assessment leads to positive consequences for those being assessed – for example, comparatively less stress than a different kind of assessment); and practicality (whether the assessment can be administered efficiently).

The survey statements were written to reflect and tap into different facets of the construct. For example, Statements 4A and 4B ("provides a meaningful measure of how students might use the target language in 'non-test' situations in the real world beyond the classroom") reflected Bachman and Palmer's (1996) authenticity argument that performance on the assessment needs to demonstrate clear resemblance to the target language use (TLU) domains targeted in the assessment. There were four sub-constructs:

1. perceived validity and reliability
2. perceived authenticity and interactiveness
3. perceived impact
4. perceived practicality.

To elicit more precise and nuanced attitudinal data than those that might have been gathered from a blunt five-point Likert scale, respondents were asked to indicate their level of response to each of the paired statements by marking a clear vertical line at the appropriate point, with "strongly disagree" at one end, and "strongly agree" at the other. The length of the response line was 5cm, and the level of response was determined by measuring the distance, in mm, from the left-hand line to the point of intersection, and then converting this into a score out of 10.

To pilot the data collection instrument, 10 teachers were invited to complete the survey and to comment on their understanding of the statements (both sections), and length of time it took them to respond to both sections of the survey.

The piloting indicated that, as an overall measure of test usefulness, all 10 items in Section I demonstrated acceptably high levels of internal consistency, whether pertaining to *converse* ($\alpha = .86$) or to *interact* ($\alpha = .73$). The "perceived validity and reliability" subscales consisted of three parallel statements – 1, 2 and 3 ($\alpha = .78$ and .90), as did the "perceived authenticity and interactiveness" sub-scales – 4, 6 and 7 ($\alpha = .86$ and .89). The "perceived impact" subscales had two parallel statements – 5 and 8 ($\alpha = .47$ and .40), as did the "perceived practicality" subscales – 9 and 10 ($\alpha = -.42$ and .23). (Statements 5 and 10 were presented in a way that reversed the polarity of the response.) With regard to perceived impact, the lower Cronbach's α scores were considered acceptable because the scales only contained two items and the average correlation between the responses for this sub-construct was $r = .45$. There was no evidence to suggest that the reversed polarity in Statement 5 was having an adverse effect. Coefficients for the "perceived practicality" subscales indicated that the two statements were not measuring this construct in an internally consistent way. Inspection of the pilot data, alongside the open-ended comments recorded in Section II, suggested that Statement 10 was not being consistently understood. Modifications to statement wording were made to the final survey.

Table 1. Numbers of survey respondents considered against numbers of NCEA (senior secondary) students.

Principal language taught	Survey respondents		Senior secondary students	
	n	%	*n*	%
Chinese	6	4	974	7
French	59	39	5011	36
German	11	7	1378	10
Japanese	42	28	3478	25
Spanish	29	19	2932	21
Unspecified	5	3	–	–
Total	152	100	13773	100

Participants

Surveys were distributed by mail to teachers across the country with a postage paid envelope for their reply. School addresses were obtained from a publicly accessible database published by New Zealand's Ministry of Education. School names were cross-checked against publicly accessible Ministry data on the languages taught. Because it was not possible to determine how many teachers of a given language taught in each school, it was recognized that not all teachers teaching a language would receive the survey (i.e., only one survey per language per school was sent out, addressed to the teacher in charge of a particular language).

A total of 579 surveys were distributed. To ensure anonymity, no names were required. The only demographic information sought was the principal language taught by the respondent, and whether or not the respondent was using *interact* at NCEA levels 1 and/or 2 at the time of the survey. 152 surveys were returned. This was regarded as a very positive response rate of just over one in four targeted FL teachers in New Zealand. When response rates across the five languages were compared to the numbers of senior secondary students (NCEA levels 1 to 3) taking each FL in 2012 (Education Counts, 2012), the response numbers correlated virtually perfectly ($r = .996$, $p < .001$), suggesting that the larger populations of teachers of these languages were adequately represented (Table 1).

Findings – Section I

The primary analysis of Section I data relates to differences in means between each of the paired statements. Scores from 0 to 10 were standardized to scores out of 100, and mean differences calculated (Table 2).

It was found that, for Questions 1 to 4 and 6 to 8, differences in the mean average +22, and the mean response is significantly higher for *interact* at $\alpha = .05$. For Questions 9 and 10, differences average −45, and the mean response is significantly higher for *converse*. On Question 5 the mean scores are not significantly different.

These data suggest that, in comparison with the one-off conversation, teachers perceived the assessments that were linked to *interact* as significantly more valid and

Table 2. Mean differences between *converse* and *interact* on each of the 10 statements.

Q	Converse			Interact			Difference			t^b	d^c	*r*	*p*
	M	*SEM*	n^a	*M*	*SEM*	*n*	*M*	*SEM*	*n*				
1	47.97	2.01	147	62.07	1.75	149	14.15	2.70	144	4.79	.801	.372	.000
2	39.71	1.95	148	62.17	1.78	150	22.24	2.76	146	7.53	1.251	.530	.000
3	42.45	2.05	148	63.04	1.84	151	20.88	2.88	147	6.88	1.139	.495	.000
4	29.93	1.92	148	59.01	2.06	150	29.08	3.02	146	8.99	1.493	.598	.000
5	32.50	2.00	146	38.85	1.99	144	5.96	2.82	139	1.92	0.327	.161	.056
6	36.53	2.00	147	63.37	1.85	150	26.99	2.84	146	8.9	1.478	.594	.000
7	29.37	1.91	147	56.69	1.97	150	27.81	2.80	146	9.26	1.538	.610	.000
8	45.60	1.91	141	57.80	1.78	141	12.17	2.59	136	4.17	0.718	.338	.000
9	73.19	1.63	146	24.00	1.76	150	−49.32	2.82	145	−17.19	2.865	.820	.000
10	68.01	1.84	146	27.42	1.98	150	−40.79	3.12	145	−12.81	2.135	.730	.000

Notes.
[a]Differences in *n* reflect missing responses.
[b]t = test for no difference between *converse* and *interact*.
[c]*d* = Cohen's *d* (no. of standard deviations in the difference).

authentic representations of students' spoken proficiency. Positive impact on the students, in terms of their likely perception of the value of each assessment as a "good test", was also significantly higher for *interact* than *converse*. However, teachers perceived *interact* as significantly less practical to administer, and the average difference in the means signals that, with regard to perceptions of practicality, *converse* and *interact* were rated as highly different in this respect. With regard to whether the assessment would be stressful for candidates (impact), neither assessment was perceived as being significantly better or worse.

Differences across languages

Figure 1 shows the mean difference score for each sub-construct displayed as a horizontal line together with the mean difference score according to language.

Prima facie the pattern of responses suggests differences in perception across languages. On average, teachers of Chinese perceived less difference between the two assessments than teachers of other languages. Teachers of German, by contrast, perceived the greatest differences, positive in terms of validity and authenticity, and negative in terms of practicality. However, the findings of one-way ANOVAs for each of the four sub-constructs indicated that none of the differences was statistically significant. When principal language taught is taken into account, there are no significant differences in perceptions.

Differences across whether or not using the assessment

Figure 2 shows the mean difference score for each construct displayed as a horizontal line together with the mean difference score according to whether or not the respondent

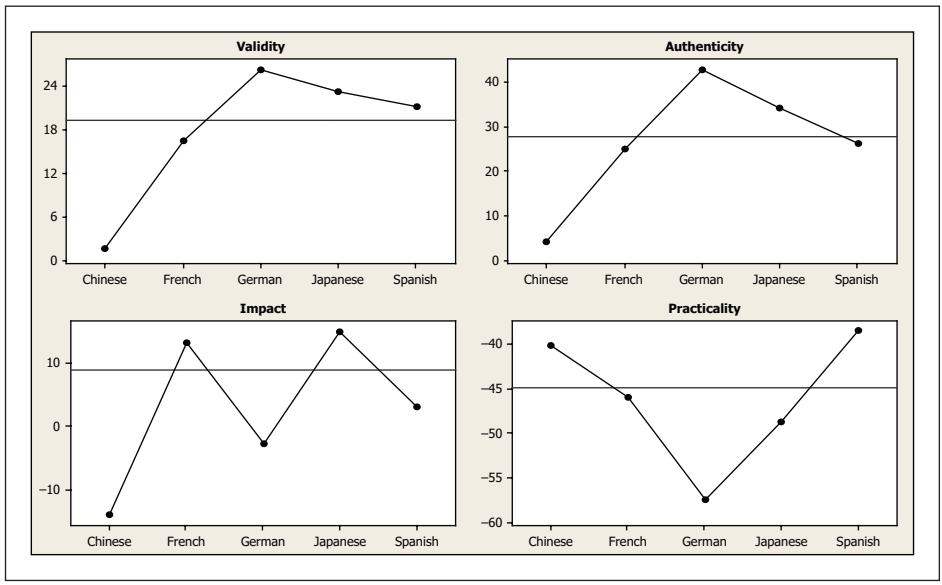


Figure 1. Sub-construct differences in mean (*converse* v. *interact*) by language taught.
Note: Panels do not have the same *y*-scale.

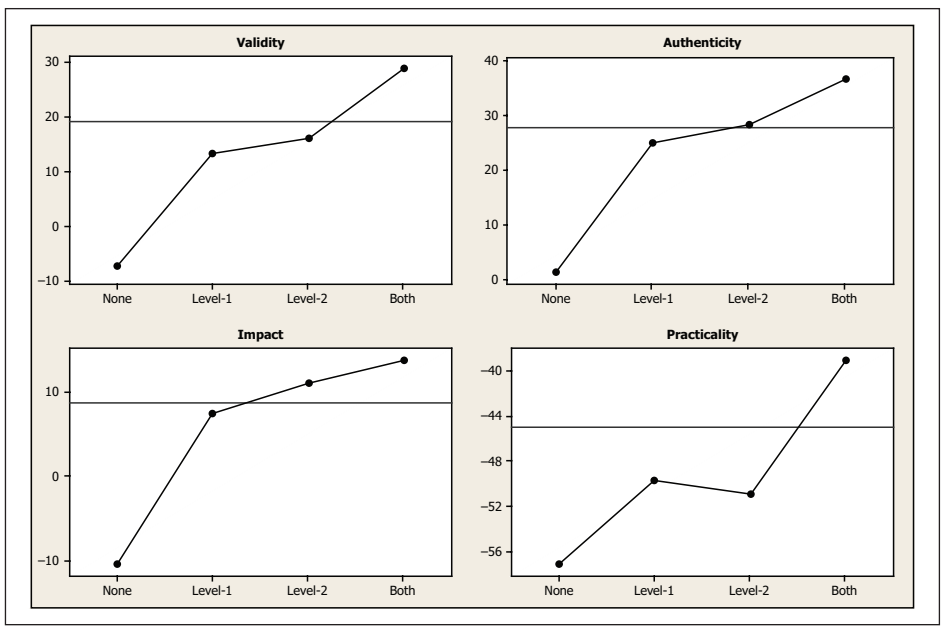


Figure 2. Sub-construct differences in mean (*converse* v. *interact*) by whether or not using *interact*.
Note: Panels do not have the same *y*-scale.

reported using the assessment at the time of the survey, whether just at level 1, just at level 2, or at both levels.

It appeared that, in comparison with those who reported not using *interact*, respondents using the new assessment not only rated it more highly than *converse* in terms of validity, authenticity and impact, but also rated it better (or, more accurately, not as badly) in terms of practicality. That is, those using *interact* appeared to perceive its benefits over *converse* more strongly, and judged its cost in terms of practicality less severely, than those who were not.

ANOVAs were conducted to compare the effect of use or non-use of *interact* on reported strengths of perception across the four sub-constructs (Table 3). A Bonferroni correction was applied, resulting in an alpha level of .0125. The differences between the two groups were highly significant for all but one sub-construct. That is, in each sub-construct other than practicality, those using *interact* rated its improvements on *converse* significantly more highly than those who were not using it. However, the assessment was perceived to be comparatively impractical, whether being used or not.

The assumption of homogeneity of variance was not violated for any of the sub-constructs (Levene's test). The assumption of normality was violated (Anderson-Darling test). This was not considered problematic, however, not only because the ANOVA is robust against the normality assumption (Hubbard, 1978) but also because, in all but one case, the differences in perception were unequivocal ($p < .001$). In the case of practicality, data were transformed for normality and a comparable result was obtained ($F = 5.49$, $p = .02$).

Findings – Section II

It is beyond the scope of this article to present an exhaustive description of the findings from Section II. What follows is a presentation of several illustrative comments related to the perceived usefulness of *interact* in comparison with *converse*, grouped according to perceived validity and authenticity (sub-constructs 1 and 2), and perceived impact and practicality (sub-constructs 3 and 4). Quotations from Section II of the survey have been cleaned (e.g., minor spelling errors corrected) to enhance readability.

Perceived validity and authenticity

High among comments related to the comparative positive advantages of *interact* was a perception that *interact* promoted authentic and valid assessment opportunities. The assessment was "refreshingly authentic" (Teacher of French, Survey 8 – French 008). Assessment tasks could be "more 'real life'" (German 057), "more genuine" (Japanese 076) and "better preparation for 'real-life' interaction in the language" (Spanish 123). Consequently, the assessments provide "a more accurate measure of the student's ability to respond to an interaction in a real-life situation" (Spanish 059), and "a genuine reflection of what they can do" (Japanese 048) and "better assessment data by collecting more evidence than conversation" (Japanese 101).

Two dimensions underscored an understanding of a broader construct of communicative spoken proficiency than the construct that appeared to be measured in *converse*: a

Table 3. Analyses of variance of difference scores for each sub-construct by use of *interact*.

Sub-construct	Source	*DF*	*SS*	*MS*	*F*	*p*
Validity	Using	1	24505	24505	30.28	.000
	Error	144	116551	809		
	Total	145	141057			
Authenticity	Using	1	24727	24727	28.28	.000
	Error	144	125901	874		
	Total	145	150628			
Impact	Using	1	10203	10203	15.51	.000
	Error	137	90155	658		
	Total	138	100358			
Practicality	Using	1	5493	5493	5.58	.020
	Error	143	140789	985		
	Total	144	146283			

focus on fluency rather than accuracy, and the opportunity to display a wider range of competencies. With "less emphasis on 'correctness' and more on communicating the message", students are "forced to use and develop the ability to respond, react, interact, and engage with the person/people they're talking to" (French 034). In other words, "the push towards more authentic, real users of the language … enables students to focus on the fillers, exclamations and sociable responses that 'oil' good conversations" (Spanish 075). Because the "focus is on interaction – that is, fillers, questions, comments, agreements etc. rather than correct structures" (French 091), students "learn to use 'natural language' – pausing, recovery, learning how to deal with [the] unexpected and not understanding" (Spanish 123).

Another perceived advantage was positive washback. The assessment was "forcing the teacher to teach in a manner which encourages communication in authentic situations" (German/French 049), "encourages the use of the target language in the classroom" (French 093) and "encourages teachers to provide many more 'authentic' speaking opportunities to students" (French 137). As a consequence, "I am doing way more speaking in the class. Interactions happen all the time whether recorded or not" (Japanese 019).

Perceived impact and practicality

With regard to perceived impact on the candidates, a mixed picture emerges. There were several dimensions in which respondents perceived that *interact* had positive impact: lowering student anxiety, multiple opportunities for assessment, increase in risk-taking and creativity.

For several respondents, lowered anxiety was directly related to the peer-to-peer nature of the assessment. That is, the fact that "students are able to converse with other students", making the assessment "less nerve wracking" (Spanish 073), was a sentiment expressed across a range of languages. Consequently students "enjoy being creative and coming up with ideas that they are going to talk about" (French 041) and are "more

willing to try things out and ask questions of each other" (Spanish 151). The assessment thereby "encourages students to take more risks when speaking" and "allows students to have much more fun" (French 013).

The assessment was also perceived as less stressful because students "are assessed over multiple occasions" (German 116) and are thereby offered "more chances … to succeed" (Japanese/French 100). That is, the assessment "takes away the 'one chance assessment'" (Spanish 024) and eliminates the "'having a bad day' reason for non-achievement" (Japanese 062), and students can "show their progress over time, rather than one single conversation" (Japanese 088). Furthermore, multiple samples are "theoretically more representative of the student output" (Spanish 021) and "give a better picture" of ability (Spanish 112).

Examples of perceived negative impact related to the fact that *interact* was still an assessment, meaning that students "want to do well" (French 013), with implications regarding the requirement to be spontaneous and unrehearsed. For some teachers, requiring spontaneity was "ridiculous" (Unstated 067), and "unrealistic" (French 132), or "too big an ask of our students" (French 008). In other words, students "find it very stressful to be put on the spot and go into a conversation unprepared" (German 114) and "cannot interact effectively without preparation" (French 041). Since *interact* is perceived as high-stakes by the students, they are "always going to want to practice beforehand, so it is no longer spontaneous" (French 093).

Interlocutor variables also contributed to perceptions of negative impact. That is, "student interactions are often dependent on their partner's ability which can make it harder for them" (French 131). This is because "sometimes the partner is not as cooperative or diligent" (Spanish 136) and "a weaker student is going to make it very difficult for a good one to show what they can really do" (French/German 141). This teacher also raised another interlocutor issue of concern: "one of my students, whose mum is German, recorded his three conversations with her, at home. This is allowed by the standard. I have a big fairness problem with this situation which angered many peers (of course who can blame the student?)."

The most stated comparative disadvantage to *interact* was practicality. The key issue related to increased workload, for teachers and students, arising from the expectation to collect at least three recorded pieces of evidence of interaction during the year. That is, increased teacher workload involved "administering it, preparing students for it, assessing it and marking each interaction, as well as managing the final portfolio of tasks" (Spanish 011). Thus, the "ongoing nature of portfolio is all consuming" (French 086), "totally stressful logistically" (Spanish 075) and "way too much work for everyone" (French 065). The classroom environment had therefore become one of "assessment driven learning" where teachers "always seem to be working on/preparing for an assessment" (Spanish 108).

Discussion

The overarching research question that the study reported here sought to answer addressed the extent to which *interact* and *converse* are supported by teacher perceptions. The quantitative data from Section I of the national survey present an unequivocal

overall teacher response: *interact* is perceived as significantly more useful than *converse except* on measures of practicality. However, both assessment types were perceived as being equally stressful for students. There were no significant differences in perception across the range of languages being assessed. The qualitative data from Section II, when considered against dimensions of usefulness, underscore the findings from the quantitative data. Seen from the perspective of teachers' thinking and beliefs (Nunan, 2004; Borg, 2006), the findings suggest the following: (1) an underlying *pedagogical* belief in the value of the communicative and interactive orientation being promoted through the revised curriculum; and (2) an underlying belief about *assessment* – that assessments, although reflecting the aims of teaching and learning, should remain distinct from teaching and learning.

Relating the first belief above to previous research into oral assessments, it is evident that several respondents appeared to value the greater authenticity promoted by *interact*, making it, in their perception, a more genuine reflection of what students could do with the language in a range of TLU domains, whether in the assessments themselves or potentially in the future (Bachman & Palmer, 1996; Galaczi, 2010; van Lier, 1989; Weir, 2005). Also, several respondents appeared to appreciate the broader construct of communicative proficiency that the assessments sought to measure (Ducasse & Brown, 2009; Jacoby & Ochs, 1995; May, 2011) alongside its positive washback (Ducasse & Brown, 2009; Messick, 1996).

However, open-ended comments serve to illustrate the potential tension between the two underlying beliefs. On the one hand, the peer-to-peer nature of the interactions and the opportunity to collect a range of evidence were appreciated, and there were several reports of students being more relaxed (Együd & Glover, 2001; Fulcher, 1996; Nakatsuhara, 2009; Ockey, 2001). On the other hand, for other teachers, student stress was still perceived as an issue – this was, after all, a high-stakes assessment and, in that respect, conferred neither advantage nor disadvantage over *converse*. In this regard, some teachers were concerned about the negative impact of interlocutor effects (Foot, 1999; Fulcher, 2003; Galaczi & ffrench, 2011; O'Sullivan, 2002), questioned whether expecting unrehearsed and spontaneous interaction was realistic and, in one case, questioned the fairness of allowing a German student to do his assessments at home. Each of these issues has implications for the perceived reliability of the assessment.

It is important to note significant differences in perception depending on whether or not the teacher had chosen to use the new assessment. In terms of validity, authenticity and impact, those using *interact* perceived the assessment more favourably than those who were not (with regard to practicality, the difference in perception was significant at $\alpha = .05$, but not at the more conservative $\alpha = .0125$). Actually using the assessment is clearly an important step in appreciating its value, although not diminishing its challenges in operationalization. It should be noted that the significant differences in perceptions here may be either *because* participants are using *interact*, or *why* they are using it.

The data have several implications when seen in the light of what the assessment reform had set out to achieve (East & Scott, 2011a, 2011b). In particular, those charged with drawing up the assessment blueprints were hoping to encourage assessments as seen from a dynamic perspective that, although contributing to a summative grade as part of a high-stakes assessment system, drew on evidence embedded in, and therefore

collected during, normal peer-to-peer interactions, whether taking place inside or outside the classroom. Teachers by contrast (and, it would seem, their students also) appear to perceive *interact* through a more traditional or "static" assessment lens, with each interaction seen as a test in its own right which should be set apart from normal classroom activity.

In other words, teachers do not yet appear to have grasped the socioculturally informed notion that students (as assessment candidates) should be allowed to "bring samples of their real language-use activities for assessment", even though teachers do support the notion of "offering a selection of tasks in formal, organised tests" (Luoma, 2004, p. 103). In turn, it is not surprising that teachers perceive *interact*, in comparison with *converse*, to be significantly more impractical (with at least three times the amount of evidence to collect, and throughout the year), that teachers do not perceive any significant difference between the two assessments in terms of candidate stress, and that some teachers report negative impact in unequivocal terms ("ridiculous", "unrealistic").

Perhaps it is necessary to recognize the impracticalities and perceptual challenges of embedding speaking assessments within normal classroom work, and to acknowledge the teachers' perception that *interact*, as a high-stakes assessment, should be operationalized as such. This would not require an abandonment of the paired assessment format, but may perhaps require a return to a static one-time assessment opportunity (a formalized test). Indeed, the increased practicality of paired assessments identified in prior studies (Ducasse & Brown, 2009; Galaczi, 2010; Ockey, 2001) presumes a one-time assessment format that is absent from *interact*. However, it must be acknowledged that the move towards a portfolio of evidence of genuine interaction arose partly from the developers' concern to address the apparent limitation that *converse* led to contrived, stilted and inauthentic interaction.

If *interact* is to remain in its current form, stakeholders need to be supported in shifting understanding towards the validity of embedding assessments seamlessly within normal classroom work, thereby moving students' attention away from the high-stakes nature of the assessment and allowing the inclusion of evidence gathered from normal day-to-day activities. Teachers would also need to embrace the notion anticipated by the assessment developers that instances of interaction outside the classroom (e.g., an interaction taking place on a trip overseas and recorded on a mobile device such as a cell phone) should be accepted as evidence. There are, however, practicality implications. For example, unless teachers effectively record all lessons with a view to perhaps catching some instance of meaningful impromptu interaction, the interactions are necessarily contrived.

These differences in operationalization (snap-shot or ongoing) raise a more fundamental issue concerning which assessment paradigm (static or dynamic) is the more appropriate to capture valid and reliable evidence of FL students' spoken communicative proficiency. East (2008) argues that neither assessment paradigm is right or wrong, better or worse. Each is "just different, and based on different assumptions about what we want to measure" (p. 9). The tension here is that *interact* may be attempting to fulfil two potentially irreconcilable functions. On the one hand, using a series of genuinely authentic interactions as evidence of spoken proficiency is intuitively appealing, and setting up stand-alone assessments for *interact* focuses on the interaction as a *test* and potentially

compromises the opportunity to collect evidence of genuine spontaneity. On the other hand, collecting lesson-embedded or "real life" evidence challenges fundamental notions of standardization and reliability that traditionally inform assessments that are used for high-stakes or accountability purposes. This tension, alongside the stakeholder voice, must inform ongoing evaluation of *interact*.

Limitations and conclusions

A key limitation of the study reported here is its reliance on teachers as a single source of evidence for test usefulness. A more comprehensive examination of comparative usefulness, and therefore claims to validity, would need to take into account the perspective of students as major stakeholders in the assessment, and also evidence derived from assessments generated under the two different conditions. In this regard, it should be noted that, in the context of the entire project, additional sources of data (not reported in this article) included interviews with teachers in 2012 ($n = 14$) and 2013 ($n = 13$), and surveys of students taking *converse* at level 3 in 2012 ($n = 30$) or *interact* at level 3 in 2013 ($n = 119$).

Another limitation is non-response bias. It is not possible to account for or explain the 74% of the sample who failed to respond. Some teachers may not have been working with students at examination level, or, if they were, may have chosen to opt out of using *interact* completely or, conversely, may have felt quite satisfied with *interact*. In each of these cases teachers may have thought it was unnecessary to respond. Others may have been too busy to respond. There is also the danger that those teachers whose underlying beliefs about effective language pedagogy may have been at odds with the emphases of the revised curriculum and assessments may have opted out of the survey.

Taking these limitations into account, the evidence gathered from this survey indicates that, at least with regard to the perspectives of the respondents, two years into the assessment reform *interact* is working relatively well. Teachers perceive *interact* to be, in most respects, a more useful assessment than *converse*. This finding provides implicit evidence that most teachers have grasped and appreciate the learner-centred and experiential nature of the new school curriculum with its emphasis, for FL programmes, on task-oriented communication and interaction (Nunan, 2004). It also provides explicit evidence that teachers see *interact* as a valid form of assessment in line with curricular aims.

More broadly, in terms of supporting the argument for the use of paired interactions as a viable alternative to single candidate interviews, the teacher evidence collected in this study provides a valuable and important stakeholder perspective. This perspective substantiates several claims for the relative beneficence of paired assessments that were outlined earlier in this article alongside several challenges for their successful implementation.

Teachers' concerns around practicality and negative impact and interaction for some candidates raise serious issues that need to be addressed. Allied to this, more thought needs to be given to the operationalization of *interact* as an assessment, in particular whether and how the tension between "real world" samples and high-stakes accountability can be reconciled, or at least work together convincingly and acceptably.

Winke's (2011, p. 633) arguments underscore the value of gathering the teacher's voice: teachers "have unique insight into the collateral effects of tests. They administer tests, know their students and can see how the testing affects them, and they recognize – sometimes even decide – how the tests affect what is taught." This "unique vantage point from which to gauge the effects of testing on students" means that teachers' perspectives are "valuable pieces of information concerning whether tests affect the curriculum as intended." In Winke's view, it is therefore "surprising" that teachers' perspectives are often not included in reviewing the validity of assessments. This study has sought to fill this gap.

Funding

This work was supported by The University of Auckland (ECREA award number 3701329).

References

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

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, UK: Oxford University Press.

Borg, S. (2006). *Teacher cognition and language education: Research and practice*. New York and London: Continuum.

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing*, *26*(3), 341–366.

Csépes, I. (2002). Is testing speaking in pairs disadvantageous for students? Effects on oral test scores. *novELTy*, *9*(1), 22–45.

Ducasse, A., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, *26*(3), 423–443.

East, M. (2008). *Dictionary use in foreign language writing exams: Impact and implications*. Amsterdam and Philadelphia, PA: John Benjamins.

East, M., & Scott, A. (2011a). Assessing the foreign language proficiency of high school students in New Zealand: From the traditional to the innovative. *Language Assessment Quarterly*, *8*(2), 179–189.

East, M., & Scott, A. (2011b). Working for positive washback: The standards-curriculum alignment project for Learning Languages. *Assessment Matters*, *3*, 94–115.

Education Counts. (2012). Subject enrolment. Retrieved 3 June 2013, from www.educationcounts.govt.nz/statistics/schooling/july_school_roll_returns/6052.

Együd, G., & Glover, P. (2001). Oral testing in pairs: A secondary school perspective. *ELT Journal*, *55*(1), 70–76.

Foot, M. C. (1999). Relaxing in pairs. *ELT Journal*, *53*(1), 36–41.

Fulcher, G. (1996). Testing tasks: Issues in task design and the group oral. *Language Testing*, *13*(1), 23–51.

Fulcher, G. (2003). *Testing second language speaking*. Harlow, UK: Pearson.

Galaczi, E. D. (2010). Paired speaking tests: An approach grounded in theory and practice. In J. Mader & Z. Ürkün (Eds.), *Recent approaches to teaching and assessing speaking*. IATEFL TEA SIG conference proceedings. Canterbury, UK: IATEFL Publications.

Galaczi, E. D., & ffrench, A. (2011). Context validity. In L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking* (pp. 112–170). Cambridge, UK: Cambridge University Press.

Hubbard, R. (1978). The probable consequences of violating the normality assumption in parametric statistical analysis. *Area*, *10*(5), 393–398.
Jacoby, S., & Ochs, E. (1995). Co-construction: An introduction. *Research on Language and Social Interaction*, *28*(3), 171–183.
Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal*, *70*(4), 366–372.
Luoma, S. (2004). *Assessing speaking*. Cambridge, UK: Cambridge University Press.
May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing*, *26*(3), 397–422.
May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly*, *8*(2), 127–145.
McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell.
Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241–256.
Ministry of Education. (2007). *The New Zealand Curriculum*. Wellington, NZ: Learning Media.
Nakatsuhara, F. (2009). *Conversational styles in group oral tests: How is the conversation constructed?*, unpublished doctoral dissertation, University of Essex, UK.
Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing*, *28*(4), 483–508.
Norton, J. (2005). The paired format in the Cambridge Speaking Tests. *ELT Journal*, *59*(4), 287–297.
Nunan, D. (2004). *Task-based language teaching*. Cambridge, UK: Cambridge University Press.
NZQA. (2009). Understanding NCEA. Retrieved 21 April 2010, from www.nzqa.govt.nz/publications/docs/understand-ncea.pdf.
NZQA. (2014a). Assessment and moderation best practice workshops. Retrieved 22 January 2014, from www.nzqa.govt.nz/about-us/events/best-practice-workshops/
NZQA. (2014b). NCEA subject resources. Retrieved 22 January 2014, from www.nzqa.govt.nz/qualifications-standards/qualifications/ncea/subjects/
NZQA. (2014c). SecQual (archive) - national moderator reports. Retrieved 20 January 2014, from www.nzqa.govt.nz/about-us/publications/newsletters-and-circulars/secqual/national-moderator-reports/
O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing*, *19*(3), 277–295.
Ockey, G. J. (2001). Is the oral interview superior to the group oral? *Working Papers, International University of Japan*, *11*, 22–40.
Rahimpoura, M. (2010). Current trends on syllabus design in foreign language instruction. *Procedia Social and Behavioral Sciences*, *2*, 1660–1664.
Taylor, L. (2001). The paired speaking test format: Recent studies. *Research Notes*, *6*, 15–17.
TKI. (2014a). New Zealand curriculum guides - senior secondary. Retrieved 25 January 2014, from http://seniorsecondary.tki.org.nz/
TKI. (2014b). Resources for internally assessed achievement standards. Retrieved 25 January 2014, from http://ncea.tki.org.nz/Resources-for-Internally-Assessed-Achievement-Standards
Turner, J. (1998). Assessing speaking. *Annual Review of Applied Linguistics*, *18*, 192–207.
van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly*, *23*, 489–508.
Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke, UK: Palgrave Macmillan.
Winke, P. (2011). Evaluating the validity of a high-stakes ESL test: Why teachers' perceptions matter. *TESOL Quarterly*, *45*(4), 628–660.

Yoffe, L. (1997). An overview of the ACTFL proficiency interview: A test of speaking ability. *Shiken: JALT Testing & Evaluation SIG Newsletter*, *1*(2), 2–13.

Appendix A: Achievement criteria for *interact*

(Searchable from www.nzqa.govt.nz/qualifications-standards/qualifications/ncea/subjects/)

Level 1: Interact using spoken target language to communicate personal information, ideas and opinions in different situations (approximate level of language: Common European Framework of Reference [CEFR] level A2 [Council of Europe, 2002])

Achievement	Achievement with Merit	Achievement with Excellence
Interact using spoken target language to communicate personal information, ideas and opinions in different situations.	Interact using convincing spoken target language to communicate personal information, ideas and opinions in different situations.	Interact using effective spoken target language to communicate personal information, ideas and opinions in different situations.

Level 2: Interact using spoken target language to share information and justify ideas and opinions in different situations (approximate level of language: CEFR level A2-B1)

Achievement	Achievement with Merit	Achievement with Excellence
Interact using spoken target language to share information and justify ideas and opinions in different situations.	Interact using convincing spoken target language to share information and justify ideas and opinions in different situations.	Interact using effective spoken target language to share information and justify ideas and opinions in different situations.

Level 3: Interact clearly using spoken target language to explore and justify varied ideas and perspectives in different situations (approximate level of language: CEFR level B1-B2)

Achievement	Achievement with Merit	Achievement with Excellence
Interact clearly using spoken target language to explore and justify varied ideas and perspectives in different situations.	Interact clearly using convincing spoken target language to explore and justify varied ideas and perspectives in different situations.	Interact clearly using effective spoken target language to explore and justify varied ideas and perspectives in different situations.

Key words such as 'clearly', 'convincing' and 'effective' are defined in the standards.

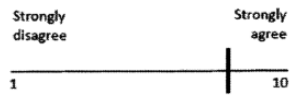
Council of Europe. (2001). *Common European Framework of Reference for languages*. Cambridge, UK: Cambridge University Press.

Appendix B: The questionnaire (key elements)

SECTION 1 – 5 minutes

The following statements compare the **new NCEA 'interact' standard (levels 1 and 2)** with the old 'converse' standard which it has now replaced. Please indicate how much you agree or disagree with each statement (there are **twenty** statements to respond to).

Please **mark a clear vertical line to indicate the level of your response**, like this:



Please do not spend too long responding to each statement. Your immediate impression is what counts.

1A	The old 'converse' standard enabled students to demonstrate clearly what they know and can do when speaking the target language	Strongly disagree … Strongly agree 1 … 10
1B	The new 'interact' standard enables students to demonstrate clearly what they know and can do when speaking the target language	Strongly disagree … Strongly agree 1 … 10

Remaining paired statements (2A/B to 10A/B) are reproduced here for 'interact' only. The new 'interact' standard:

2. provides an accurate measure of students' spoken communicative proficiency
3. provides good opportunities to measure students' fluency in the target language
4. provides a meaningful measure of how students might use the target language in 'non-test' situations in the real world beyond the classroom
5. Students generally find that completing the new 'interact' standard makes them feel anxious and stressed
6. promotes the opportunity for students to engage in genuine social interactions in the target language
7. promotes the opportunity for students to use authentic and unrehearsed language
8. Students generally feel that the new 'interact' standard is a good test of their spoken ability
9. is easy to manage and administer
10. To administer the new 'interact' standard takes up a lot of class time at the expense of the available teaching time.

SECTION 2 – 10 minutes

1. What do you see as the main *advantages* of the new 'interact' standard in comparison with the old 'converse' standard?
2. What do you see as the main *disadvantages* of the new 'interact' standard in comparison with the old 'converse' standard?
3. **EITHER:** Briefly describe your *experiences* with introducing the new 'interact' standard at levels 1 & 2. How did it go? What problems (if any) did you encounter? How did the students respond?
   OR: Briefly explain why you decided NOT to use the new 'interact' standard
4. Any advice on how the new 'interact' standard might be improved?