Addendum

Field Testing the OPSBA Speaking Assessment: Contributors, Process, Results and Psychometric Analysis

Field-testing an assessment before implementation is considered good practice and is especially important when high stakes decisions are going to be made based upon the results of that assessment. This addendum supports the *French-Language Proficiency Assessment for French as a Second Language Teaching: A Toolkit and Resource Guide, May 2023***.** It contains:

I. Names of the validation team members who provided feedback and support of the field-testing process;
II. Field-testing locations;
III. Assessment development process, and
IV. Results from the analyses of the data collected during field testing of the OPSBA speaking assessment.

The purpose of the field test was to gather evidence that the assessment targeted appropriate language skills, was at an appropriate level of difficulty, had acceptable reliability, and could be easily implemented by organizations such as school districts. Here, we present evidence to support the validity, reliability, and feasibility of the OPSBA speaking assessment. For readers who have developed their own assessment (or are considering developing one), reading this document will also give some ideas about the types of analyses that can be used to evaluate the quality of their own assessment.

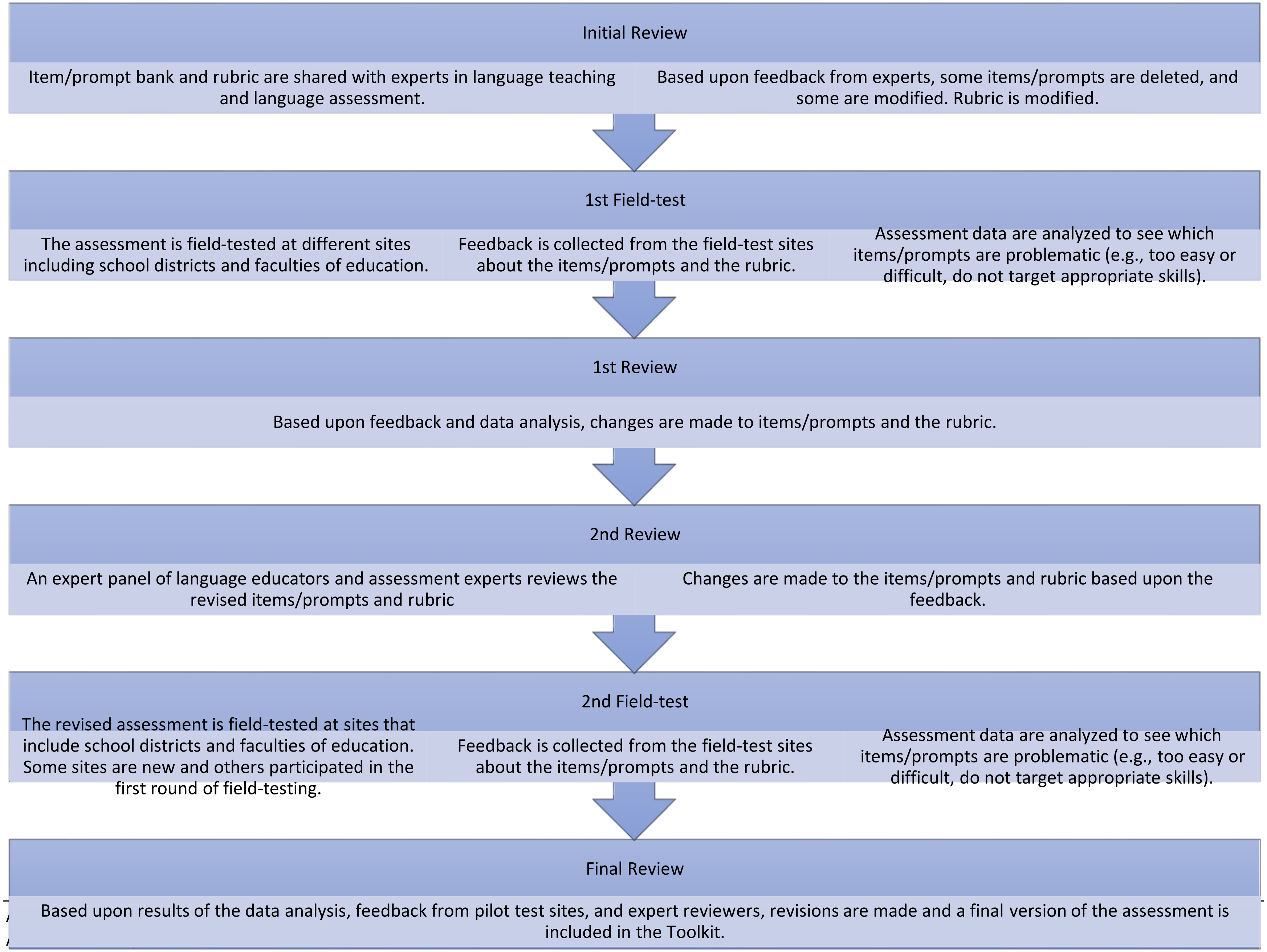
Table 1. Contributors

Validation Team Members	Field-test sites for OPSBA assessment
Léo-James Lévesque, St. Thomas University	Delta School District, British Columbia
Dr. Jennifer Straub, Wilfrid Laurier University	Queen's University
Dr. Mimi Masson, Université de Sherbrooke	Renfrew County District School Board
Shelley Gagné, Renfrew County DSB	St. Thomas University
Dr. Katherine Mueller, University of Calgary	Toronto Catholic District School Board
	Tyndale University
Pierre Riopel, President (Ret.), Collège Boréal, Sudbury	University of Alberta, Faculté St Jean
Janine Griffore, Assistant Deputy Minister (Ret.), Ontario Ministry of Education	Vancouver School Board
	Waterloo Catholic District School Board

Danielle Couture, Educator Certification and Learning Management Coordinator, Government of Northwest Territories
Catherine Youngblud, Principal (Ret.), Hamilton-Wentworth DSB

Figure A

Assessment Development Process



Initial Review

Item/prompt bank and rubric are shared with experts in language teaching and language assessment.

Based upon feedback from experts, some items/prompts are deleted, and some are modified. Rubric is modified.

1st Field-test

The assessment is field-tested at different sites including school districts and faculties of education.

Feedback is collected from the field-test sites about the items/prompts and the rubric.

Assessment data are analyzed to see which items/prompts are problematic (e.g., too easy or difficult, do not target appropriate skills).

1st Review

Based upon feedback and data analysis, changes are made to items/prompts and the rubric.

2nd Review

An expert panel of language educators and assessment experts reviews the revised items/prompts and rubric

Changes are made to the items/prompts and rubric based upon the feedback.

2nd Field-test

The revised assessment is field-tested at sites that include school districts and faculties of education. Some sites are new and others participated in the first round of field-testing.

Feedback is collected from the field-test sites about the items/prompts and the rubric.

Assessment data are analyzed to see which items/prompts are problematic (e.g., too easy or difficult, do not target appropriate skills).

Final Review

Based upon results of the data analysis, feedback from pilot test sites, and expert reviewers, revisions are made and a final version of the assessment is included in the Toolkit.

Results from the analyses of the data collected during field testing of the OPSBA speaking assessment

The first part of the *Addendum* summarizes the qualitative data collected from our expert reviewers and from the assessment users at our field test sites. The second part presents the results of the psychometric analyses performed on the quantitative data.

Qualitative Data

Feedback from the first round of pilot testing

Feedback from the first round of pilot testing indicated the items/prompts addressed language skills that were required of FSL teachers, but the rubric was complicated and cumbersome to use. Table 2 summarizes the feedback we received from expert reviewers and field test sites along with the action taken.

Table 2. Qualitative feedback on first draft of OPSBA Speaking Assessment

Feedback	Action Taken
Items/Prompts are appropriate and reflect language skills that would be used by FSL teachers	None
Need better recording quality on the item/prompt that uses a sample of French spoken by a FSL learner	Rerecord the audio sample so it is clearer
Need a selection of images for the item which used an image to elicit a response	A selection of images was included based upon the recommendations of the second expert panel
Simplify the rubric	The original two-part rubric was slimmed to a single rubric. The number of scoring criteria was reduced from 12 to 5.
Ensure scoring criteria match the language sample elicited by the prompts	Scoring criteria were modified to better reflect the language sample elicited by the conversation prompts.

Feedback from the second round of field testing

The second expert panel agreed that the items would elicit speaking skills and vocabulary germane to FSL teaching. There was also agreement that the scoring criteria in the rubric covered a range of language skills used in FSL teaching and that descriptions of each achievement level were clear. Some individuals on the panel identified aspects of the assessment they thought could be improved, but there was no consensus that any aspect of the speaking assessment was problematic. As an example, one item/prompt that used an image of a painting of the voyageurs generated much discussion about how to best incorporate Indigenous content into the assessment. There was little consensus among panelists about how to do this and so we asked different Indigenous educators for their perspective on the item and images used as conversation prompts. Once again, there was no consensus and so we used the expertise of a language education specialist who identified as Métis to make a final decision about what images to include and how to phrase the question in this item/prompt.

Feedback from the field test sites indicated that the new rubric was much easier to use, which, in turn, enhanced the reliability of the assessment. One field test site recorded all the conversations and found that using recordings allowed the assessors to focus on the conversation and engaging the candidate. Because the conversation was recorded, assessors could conduct the scoring at a time that was convenient to them and made it possible to re-listen to the candidate's responses multiple times, enhancing their ability to detect errors and strengths in the candidate's speaking abilities. Scoring in this way also ensures that scorers come to independent evaluations of the candidate.

Finally, all field test sites spoke about the value of assessors talking about the rubric and items/prompts before conducting the assessment. These conversations allowed the assessors to become more familiar with the rubric, decide which prompts would be most suitable for their purposes, and determine exactly how the assessment would be implemented. Field test sites also talked about the importance of assessors coming together after the assessments to discuss the scores they awarded. This process allowed assessors to better define what separates different levels of achievement and become more consistent in their scoring.

Quantitative data

First round of field tests

We present a very brief summary of the results of the first round of field testing. This is because the quantitative data from the first round of field testing was inconsistent in its quality and so not all data were useful for statistical analyses. The quantity of missing data, along with the sample size (N = 63 candidates) prevented the use of sophisticated psychometric analyses. Nonetheless, the data indicated that that the rubric and items were pitched at an appropriate level of difficulty as the scores tended to be in the middle of the range instead of at the extremes. Mean scores on each criterion (out of 4) ranged from 2.40 (use of agreements. e.g., subject – verb; noun – adjective; noun – article) to 3.11 (alters language to clarify and/or illustrate complex ideas).

One important finding from the quantitative data was that many times one scorer would assign a score to a criterion while another rater would assign a value of "NA" meaning the scorer was not able to assess the criterion due to a lack of evidence. This situation arose because of the complexity of the rubric. The cognitive load of paying attention to the candidate's/applicant's responses while scoring multiple criteria at the same time meant that scorers were not able to reliably assess every criterion. This finding, along with the qualitative feedback we received, spurred the drive to simplify the rubric for the second round of field testing.

Second round of field tests

For the second round of field tests more prescriptive instructions were given to test sites about how to conduct the field tests. This round of field testing had a slightly larger sample size (N = 95) and higher quality data, allowing a broader range of statistical analyses to be performed.

Item/Prompt Difficulty

The first analysis was to examine the difficulty of each criterion (Table 3). As can be seen, the "Use of pace and inflection to facilitate understanding" criterion yielded higher scores than the other 4 criteria, but a one-way ANOVA found no statistically significant differences in mean score across the five criteria ($F(4) = 2.26$, $p = 0.06$).

Table 3. Mean scores on speaking criteria

Criterion	**Mean Score**	**Standard Error**
Use of a range of verb tenses to facilitate communication	3.01	0.09
Agreement Knowledge and use of gendered nouns in French, and related agreements	2.89	0.09
Vocabulary is accurate, demonstrates breadth, and facilitates communication	3.01	0.08
Knowledge and use of French syntax	2.92	0.09
Use of pace and inflection to facilitate understanding	3.22	0.08

It should be noted that at one field test site all of the people assessed (except one) were francophone and scored very high on the assessment. This is likely why the mean scores in the second round of field testing were higher than in the first round.

Internal structure

To examine the internal structure of the OPSBA Speaking Assessment, exploratory factor analysis was used. Although the sample size is small for factor analysis, factor analysis can yield interpretable results from small samples provided the number of factors is also small (i.e., one or two factors). Both the Kaiser-Meyer-Olkin test (*KMO* = 0.90) and Bartlett's test of sphericity ($p < .001$) indicated the data were likely to yield interpretable results from the factor analysis and this proved to be the case. Using a maximum likelihood extraction algorithm we found a one-factor model best fit the data. This was evidenced both by the scree plot and by the Kaiser criterion. The single factor accounted for 78.6% of the variance in scores. All five criteria loaded strongly onto the single factor (Table 4).

Table 4. Factor loadings of the five criteria.

Criterion	Factor Loading
Use of a range of verb tenses to facilitate communication	0.87
Agreement Knowledge and use of gendered nouns in French, and related agreements	0.80
Vocabulary is accurate, demonstrates breadth, and facilitates communication	0.92
Knowledge and use of French syntax	0.90
Use of pace and inflection to facilitate understanding	0.79

The "use of pace and inflection to facilitate understanding" likely had the lowest factor loading because it is the criterion with the least variance and less variance leads to lower correlations and therefore lower factor loadings. The one-dimensional nature of the data suggests that it is psychometrically appropriate to report scores using a single number (e.g., the candidate or applicant's total score). This can make reporting and comparison easier than if multiple scores need to be reported. Of course, if the assessment is to be used for formative purposes, it is vital that all criterion scores are reported and discussed with the candidate/applicant.

The one-dimensional factor structure suggests that the five scoring criteria should have good internal consistency as measured by Cronbach's alpha and this was found to be the case (α = 0.93). This is a very high value suggesting that ratings across the five criteria tend to be similar. All criteria contributed positively to the internal consistency of the assessment.

Consistency among scorers

For reasons of reliability, validity, and defensibility, it is important that different assessors agree upon scores for candidates. Assessment results should depend solely on the candidate performance and not upon who conducted the assessment. In language assessment (especially speaking and writing), there is always some subjectivity on the part of the scorers. It is important that the assessment items/prompts, rubric, and training lead scorers to make consistent decisions about the level of a candidate's or applicant's performance. When scores are unduly influenced by factors other than candidate or applicant performance, the assessment becomes unfair, and the resulting decisions become difficult to justify.

To provide evidence for the consistency of scoring across raters we examined the results from a field test site that was a school district (n = 11) and from two faculty of education sites (n = 9 and n = 15). The school district and first faculty of education site used two scorers for every applicant/candidate and the second faculty of education site used three scorers for every candidate. As can be seen from Table 5, the consistency between the two scorers (known as interrater reliability) at the school district field test site was very high. Pearson's correlation coefficient was used as a measure of consistency between scorers but supplemented with other measures. Cohen's Kappa is a measure of exact agreement between scorers above and beyond random chance. The weighted Kappa statistic is similar but adds a linear weighting factor for extreme disagreements. Because there were no extreme disagreements in this data set, the weighted Kappa values are higher than the values for Cohen's Kappa. Kappa values (whether Cohen's or weighted) have a maximum value of 1 (perfect agreement). Negative Kappa values can exist, and this means the level of agreement was worse than random chance. Values above 0.40 are considered reasonable and those above 0.80 are considered excellent. Finally, we also looked at the percentage of ratings where the two scorers agreed exactly. As can be seen in Table 5, the two scorers had exact agreement about 80% of the time.

No single interrater reliability statistic gives a perfect measure of scorer consistency, but the four statistics displayed here demonstrate that interrater reliability at this site was excellent.

Table 5. Interrater reliability at a school district field testing site (2 scorers)

Criterion	Pearson's *r*	Cohen's Kappa	Weighted Kappa	Percentage Agreement
Use of a range of verb tenses to facilitate communication	0.90	0.73	0.80	82%
Agreement Knowledge and use of gendered nouns in French, and related agreements	0.82	0.59	0.67	73%
Vocabulary is accurate, demonstrates breadth, and facilitates communication	0.88	0.72	0.78	82%
Knowledge and use of French syntax	0.95	0.86	0.90	91%
Use of pace and inflection to facilitate understanding	0.82	0.71	0.75	82%

The two assessors at this site were experienced language assessors who had worked together as their district's language assessment team for several years. Before implementing this assessment, they familiarized themselves with the rubric, items/prompts, and had several discussions about how they were going to implement the assessment and interpret the scoring criteria. This high level of advance preparation, coupled with their prior experience as language assessors, likely enhanced their ability to make consistent decisions about the French language proficiency of applicants to their district.

At the first faculty of education site, the scorers were from professional practice and not from academia. At this site, there were four scorers, but we only used data from two scorers. The first pair of scorers agreed on every single rating of every candidate and so these data were rejected on the belief that these ratings were not genuinely independent. The second pair assessed nine candidates. The interrater reliability statistics of this second pair is shown in Table 6.

Table 6. Interrater reliability at faculty of education field testing site #1 (2 scorers).

Criterion	Pearson's r	Cohen's Kappa	Weighted Kappa	Percentage Agreement

Use of a range of verb tenses to facilitate communication	0.78	0.45	0.53	67%
Agreement Knowledge and use of gendered nouns in French, and related agreements	0.59	0.31	0.44	56%
Vocabulary is accurate, demonstrates breadth, and facilitates communication	0.27	0.20	0.23	50%
Knowledge and use of French syntax	0.25	-0.05	0.08	22%
Use of pace and inflection to facilitate understanding	0.83	0.04	0.13	44%

The interrater reliability at this field test site was lower than at the school district site. As can be seen from Table 6, the criteria related to verb conjugation and agreements had reasonable interrater reliability. The other three criteria had poor interrater reliability. We hypothesize a few reasons for this. For the vocabulary criterion, one of the raters indicated they were not able to rate one of the candidates on this criterion. This reduced the sample size from 9 to 8, reducing the robustness of the statistics. For the syntax criterion, there was one instance of an extreme disagreement (i.e., a difference of 2 in the score) which reduced the interrater reliability substantially. The pace and inflection criterion also had substantial disagreement, but we noted that one scorer was consistently lower than the other in the scores awarded. These types of consistent disagreements can often be resolved with scorer training. Conversely, scoring disagreements in the syntax criterion did not display a pattern.

The second faculty of education field testing site used three scorers instead of two, and so different statistics were used to calculate interrater reliability (Table 7). With three raters, there are three possible Pearson correlation coefficients for each criterion, so the range of correlation coefficients is reported. The last column of Table 7 shows Fleiss’ Kappa which is a statistic used when there are 3 or more raters. Fleiss’ Kappa can have values between -1 (agreement is worse than random chance) and +1 (perfect agreement). Interpreting values of Fleiss’ Kappa can be difficult, but higher positive values are better. The values for Fleiss’ Kappa were all positive, but the value for the “Use of a range of verb tenses to facilitate communication” was low. The Fleiss’ Kappa values for the other four criteria are reasonable. The results in Table 7 show that while all correlation coefficients were positive, the values were lower than for the school district field test site but more consistent than for the other faculty of education field test site.

Table 7. Interrater reliability at faculty of education field testing site #2 (3 scorers).

Criterion	**Pearson's *r***	**Fleiss' Kappa**
Use of a range of verb tenses to facilitate communication	0.50 – 0.51	0.09
Agreement Knowledge and use of gendered nouns in French, and related agreements	0.59 – 0.77	0.31
Vocabulary is accurate, demonstrates breadth, and facilitates communication	0.51 – 0.72	0.32
Knowledge and use of French syntax	0.51 – 0.84	0.33
Use of pace and inflection to facilitate understanding	0.43 – 0.74	0.44

At this field test site, one scorer worked within the francophone system, one was an assessment professor at the faculty who had never taught French, and the third had extensive experience as a teacher and administrator within French Immersion programs. The three scorers had never worked together scoring a language assessment and this likely lowered the consistency of scoring.

To determine to what extent scorer disagreements might lead to problems in decision making, we looked at the total score awarded by the scorers at each field test site. At the school district field test site, the total score awarded (out of 20) by the two scorers never disagreed by more than one point. At the first faculty of education field test site, the total score awarded was within +/- 1 of the average score for 75% of the candidates. At the second faculty of education field test site, the total score awarded was within +/- 1 of the average score for 14 of the 15 candidates.

The weight of the evidence suggests that the OPSBA developed speaking assessment leads to good consistency across scorers. The consistency is enhanced by using total scores (instead of relying on the score of a single criterion). The field test results indicate that any differences in total score are small and unlikely to lead to differential decision outcomes. The small differences also mean that coming to a consensus on a final awarded score is likely to be easy.

How many scorers and criteria are needed?

Every assessment is a trade-off between reliability and economy. Assessments can be made more reliable by using more items and more scorers, but this comes at a cost. Additional funds, time, and personnel are needed when assessments become longer and use more scorers. Given that the factor analysis and Cronbach's alpha results show the speaking assessment has very high internal consistency, it is worthwhile to investigate whether making the assessment shorter can yield reliable results in less time. Additionally, the user's guide for the assessment suggests using more than one scorer, but with decent consistency across scorers, it is possible that using the assessment with a single scorer would yield reliable, defensible results.

To investigate under what conditions the OPSBA speaking assessment would achieve acceptable levels of reliability, we used a procedure called G-theory. G-theory analyzes the sources of variance within the scores to determine which sources are most important. Ideally, the most important source of variance is the person taking the assessment, but other sources can include who is assessing and which criterion is being assessed. One advantage of G-theory is that findings can be used to inform a D-study. A D-study will tell you how having different numbers of scorers and different numbers of criteria will affect the reliability of the assessment.

The G-study incorporated a fully crossed three facet design where the candidate or applicant was one facet, the scorer was another, and the scoring criterion was the third. A generalizability coefficient of 0.90 or higher is considered an acceptable level of reliability in a high stakes assessment.[1] A coefficient of 0.80 or higher is considered acceptable in low stakes assessments. While the OPSBA Speaking Assessment is likely to be used in high stakes scenarios (e.g., hiring) it is not the only assessment used in the decision-making process thereby reducing the stakes associated with it.

The scoring data from the two faculty of education field test sites were used for the G-theory analysis because it was felt that the conditions at the school district field test site (i.e., two experienced language assessors who had worked together for years) would likely not represent the assessment conditions in most organizations.

In a G-theory analysis, the first step is to calculate variance components. The results of this analysis are in Table 8. As can be seen the candidate or applicant (labeled 'person' in the table) is the biggest source of variance (outside of the error term). This is the desired result as the person taking the test should have a bigger impact on the score than any other source of variance.

[1] Kim, Y. S. G., Schatschneider, C., Wanzek, J., Gatlin, B., & Al Otaiba, S. (2017). Writing evaluation: Rater and task effects on the reliability of writing scores for children in Grades 3 and 4. *Reading and writing*, *30*, 1287-1310.

Table 8. Variance components in candidate or applicant scores at Faculty of Education field test sites.

Source of Variance	Variance Component Site #1	Variance Component Site #2
Person	0.271	0.317
Scorer	0.032	0.010
Criterion	0.009	0.054
Person * Scorer	0.049	0.034
Person * Criterion	0.000	0.036
Scorer * Criterion	0.013	0.001
Error	0.195	0.176

Using these variance components, we calculated a generalizability coefficient of G = 0.53 for one scorer on one criterion at site #1 and G = 0.56 at site #2. The subsequent D-study demonstrated how changing the number of scorers and number of criteria affected the reliability of the test (Table 9).

Table 9. Generalizability coefficients under different assessment conditions at faculty of education field test site #2.

Number of scorers	Number of scoring criteria	Generalizability coefficient Site #1	Generalizability coefficient Site #2
1	1	0.53	0.56
1	4	0.73	0.78
1	5	0.75	0.81
2	1	0.69	0.69
2	4	0.85	0.87

2	5	0.86	0.88
3	5	0.90	0.91

As can be seen in Table 9, the more scorers there, and the more scoring criteria are included, the higher the generalizability coefficient and the greater the reliability of the assessment. The OPSBA Speaking Assessment rubric has 5 scoring criteria and faculty of education site #1 used two scorers, so their generalizability coefficient was G = 0.86 which is slightly below the G = 0.90 threshold recommended for high stakes tests. Faculty of education field test site #2 used three scorers, so their generalizability coefficient was G = 0.91 which is above the 0.90 threshold for high stakes tests. Had this site used only two scorers their generalizability coefficient would have been G = 0.88 which is only marginally below 0.90. In our view, this would not be problematic because the OPSBA speaking assessment will not be the only assessment used in making hiring or admissions decisions. Further, the two sites with the lowest consistency in scoring were used in these G-theory analyses. This means the results may be somewhat pessimistic.

It should be noted that a common practice in many school districts is to use one assessor who gives a holistic judgement of the candidate's French language proficiency. This situation is analogous to having one scorer and one criterion. In this case, the generalizability coefficients were very low (G = 0.53 and G = 0.56) meaning the reliability of such an assessment is extremely problematic. For defensibility purposes, it is critical that users of the OPSBA Speaking Assessment use more than one scorer when evaluating a candidate. As an example, faculty of education field test site #1 had scorers whose agreement was worse than random chance on the syntax criterion indicating the possibility of idiosyncratic and inconsistent ratings. However, by using multiple scorers and multiple criteria, the overall reliability of the assessment at that site became high enough to support high-stakes decision making.

Key Takeaways

No assessment is perfect, but our field testing process and results provide support for the following claims about the OPSBA Speaking Assessment.

1. It is at an appropriate level of difficulty.
2. It is a one-dimensional assessment and that dimension is related to speaking skills needed for FSL teaching.

3. Consistency among scorers is sufficient to make reliable, defensible decisions. This is especially true when the scorers familiarize themselves with the assessment before implementation and come to a consensus about how to interpret the scoring rules.
4. Using all five scoring criteria and at least two scorers enhances the reliability of the assessment and therefore the defensibility of decisions arising from the assessment results.
5. Recording the conversation allows scorers to complete their work independently and with a greater degree of confidence.