Validity of Measurement and Proficiency Level Hierarchy in the Reading and Listening sections of the STAMP 4S

Victor D. O. Santos, PhD

Director of Assessment and Research

Avant Assessment LLC

11/5/2019

*Avant STAMP 4S is a proficiency-oriented assessment of listening, reading, writing and speaking

NOTICE: The contents of this report were developed under a grant from the Department of Education. However, those contents do not necessarily represent the policy of the Department of Education, and you should not assume endorsement by the Federal Government.

Validity of Measurement and Proficiency Level Hierarchy in the Reading and Listening sections of the STAMP 4S

Victor D. O. Santos, PhD
Director of Assessment and Research
Avant Assessment
November 5, 2019

The STAMP Test

The STAMP (STAndards-based Measurement of Proficiency) test assesses test-takers' language proficiency across the four language domains: reading, writing, listening, and speaking. The Reading and Listening sections of the test are automatically scored and results are available to test-takers immediately. The Speaking and Writing sections are scored by Avant raters trained on the STAMP proficiency scale, which is based on the ACTFL Proficiency Guidelines. The Speaking and Writing results are available between 3-7 days after the section is completed.

The STAMP test has been in operation for over 20 years and was the first computer-adaptive, online test of language proficiency. Proficiency scores for each of the four sections are reported on the STAMP scale (1-9)[1], with each of the STAMP levels corresponding to one of the ACTFL proficiency sublevels, as shown below:

STAMP Level 1	STAMP Level 2	STAMP Level 3	STAMP Level 4	STAMP Level 5	STAMP Level 6	STAMP Level 7	STAMP Level 8	STAMP Level 9
Novice			Intermediate			Advanced		
Low	Mid	High	Low	Mid	High	Low	Mid	High

The STAMP delivery engine is adaptive. The difficulty of the items that a test-taker encounters during the 30-item Reading and Listening sections of the test adapts in real time to the system's statistical estimate of that test-taker's language proficiency, based on his or her responses to all questions answered up to that point. This ability of the STAMP system to adapt in real time to the estimated proficiency level of test-takers allows for a more precise measurement of language proficiency (Hendrickson 2007; Wainer, Kaplan, & Lewis, 1992; Weiss, 1985), reduces item exposure (Linacre, 2000; Thompson, 2011; Yan, Lewis, & Stocking, 2004), therefore increasing test security, and provides a more targeted, efficient, and pleasant test-taking experience (Linacre, 2000) since test-takers will not encounter many items that are

[1] Reading and Listening sections are scored up to a STAMP level of 9 (Advanced-High). Writing and Speaking sections are scored up to a STAMP level 8 (Advanced-Mid).

Copyright © 2019 Avant Assessment, LLC. All rights reserved.

substantially above or below their ability. This is in contrast to a linear or fixed form test, in which all test-takers see the exact same test items, jeopardizing test security, decreasing the precision of measurement when compared to adaptive tests, and leading to a more frustrating testing experience for many test-takers.

The adaptivity of the STAMP 4S test is also incorporated into the Writing and Speaking sections, to some extent. The three prompts that test-takers respond to in the Writing and Speaking sections are selected from a larger pool of prompts depending on their score in the Reading and Listening sections, respectively. This ensures that the prompts in the Writing and Speaking sections of STAMP are well targeted to each test-taker, increasing the chances that test-takers will be able to produce language that closely reflects their actual proficiency in these two domains.

Hierarchy of Proficiency Levels in the Reading and Listening sections of the STAMP 4S

In order to assign specific proficiency levels to test takers at the end of the Reading and Listening sections of a STAMP test, the STAMP automated system takes into account all of the questions the test-taker was presented with, the targeted proficiency level (Novice, Intermediate, or Advanced) of each of those questions, and the actual statistical difficulty of each question. Every item in a STAMP test is calibrated through a measurement framework called Item-Response Theory (IRT), which allows for items in a certain section of a STAMP test to be compared against one another in terms of difficulty and for the system to assign the appropriate proficiency level to a test-taker at the end of a section.

In order to evaluate the measurement validity of the STAMP 4S test items and proficiency scores, the following two questions must be answered:

Question 1: What is the difficulty hierarchy among Novice, Intermediate, and Advanced items in the STAMP 4S?

Question 2: Do test-takers who receive a higher STAMP score indeed have a higher level of proficiency in the language?

In the subsequent sections, we present STAMP 4S results that will allow us to answer the two critical questions above.

Copyright © 2019 Avant Assessment, LLC. All rights reserved.

Question 1: What is the difficulty hierarchy among Novice, Intermediate, and Advanced items in the STAMP 4S?

In order for a language assessment that bases its scores on the ACTFL Proficiency Guidelines (ACTFL, 2012) to be defensible, it is vital that test developers can show, based on data from a large number of real administrations of the test, that the average difficulty of Novice items on the test is lower than that of Intermediate items, which in turn should be lower, on average, than that of items written to target the Advanced level (Cox & Clifford, 2014). If this item difficulty hierarchy turns out not to hold in a test for the three major proficiency levels, this would constitute a serious threat to the validity of the test. For instance, if Novice items in a test were harder, on average, than Intermediate items, test-takers receiving a final score of Intermediate-Mid after completion of the test could in fact be less proficient than those receiving a score of Novice-High.

As many language testing practitioners will be quick to point out, it is not rare that an item is sometimes developed to target a specific proficiency level but turns out to be statistically easier or harder than intended when actual test data is analyzed. This can be due to several factors, including the quality of the item writing and experience of the item writers, mis-leveling of the item in relation to the proficiency levels, test-taking strategies being used by test-takers that allow them to correctly answer an item despite not having the level of proficiency required to understand the passage, and many other factors. When the empirical difficulty of an item is far from what test developers would expect given the proficiency level the item was written to target, test developers must have processes in place to review that item and address potential problems with it. Some of the possible actions to be taken for such an item include content revision, removal from the pool of test items and replacement with a more statistically appropriate item, reassessment of its targeted proficiency level and others.

To answer Question 1, we will look at the empirical difficulties of actual STAMP test items at the three major proficiency levels (Novice, Intermediate, and Advanced) for some of the languages in which STAMP is available. STAMP 4S items were calibrated in the statistical software Jmetrik (Meyer, 2014). The languages considered for this analysis are the following[2] (the number in parentheses indicates the number of test-takers the analysis is based on) and results are based on recent data and reflect the state of current STAMP 4S tests.

Spanish (n = 83,849)
French (n = 58,465)
Japanese (n = 6,930)
Chinese Simplified (n = 8,803)
Arabic (n = 4,915)

[2] These languages were chosen to present results for a diverse group of STAMP languages.

Copyright © 2019 Avant Assessment, LLC. All rights reserved.

Below, we examine the hierarchy of item level difficulties for the Reading and Listening sections of each of the five STAMP 4S tests above. Statistical analyses of the differences between items at the three major proficiency levels (Novice, Intermediate, and Advance) is also presented.

Spanish Reading and Listening

Figure 1 shows the average item difficulty for Novice, Intermediate, and Advanced items in the **Reading** section of STAMP 4S Spanish. The black horizontal line inside the box plot for each of the three proficiency levels represents the median[3] IRT difficulty for items at that level . The green horizontal line shown for each of the three proficiency levels indicates the average/mean item difficulty for that level, and is what we are mostly interested in. In the IRT framework used, item difficulties usually range from - 4 to + 4. Higher numbers on the y-axis (Rasch Difficulty) indicate more difficult items or groups of items. Any dots above or below a box plot indicate outlier items, which are significantly easier or harder than other items at the same major proficiency level.

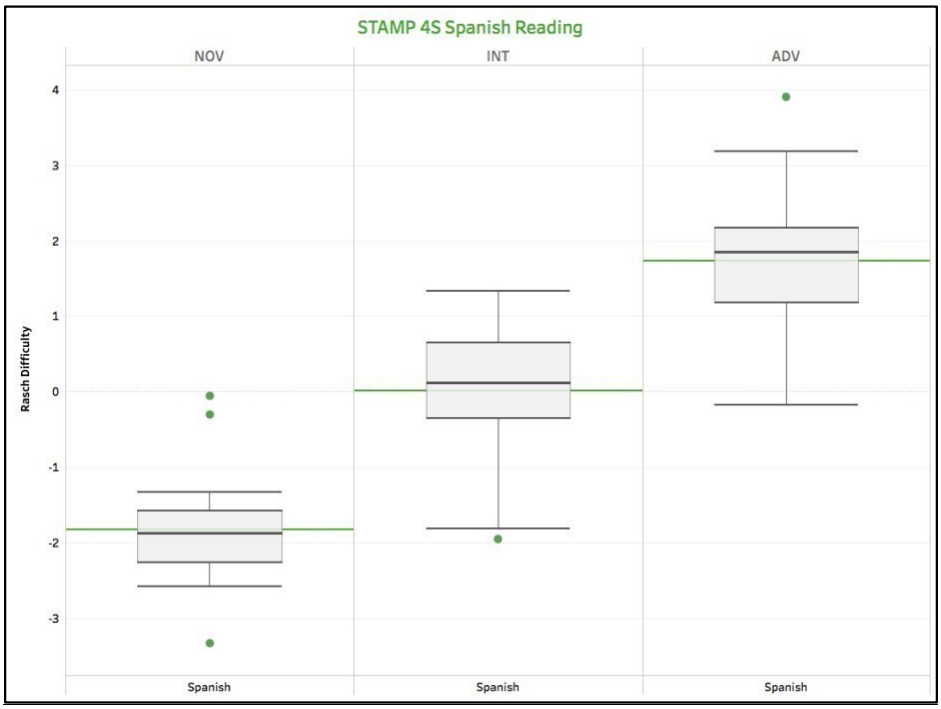


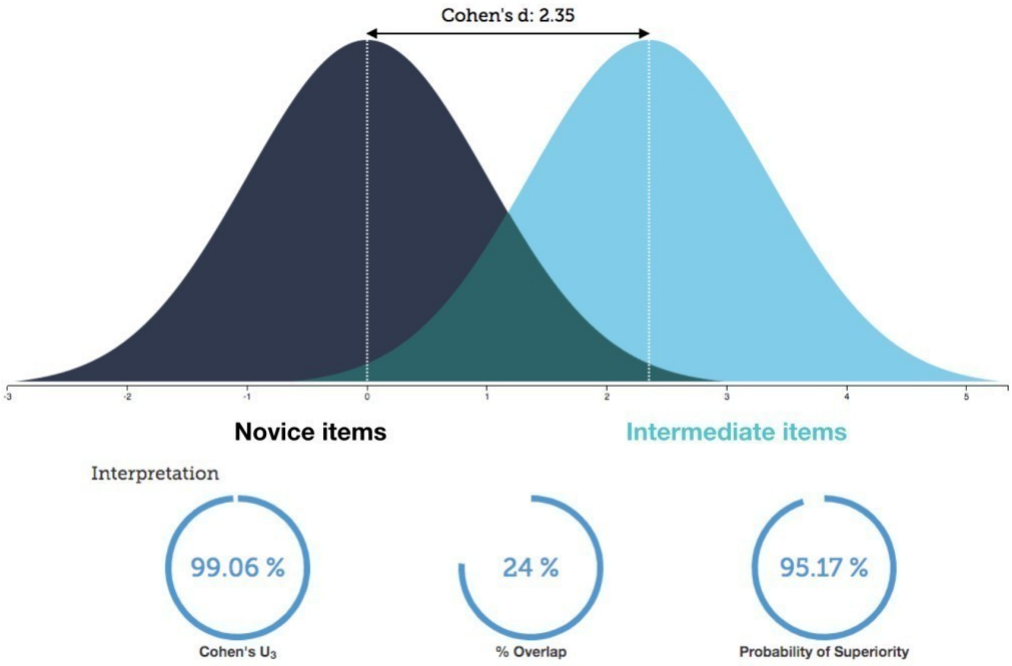
Figure 1. Mean item difficulty of Novice, Intermediate, and Advanced items in STAMP 4S Spanish (Reading)

[3] The median difficulty for items at a level indicates that 50% of the items at that level have a difficulty above the median value and 50% of the items have a difficulty below the median value.

Copyright © 2019 Avant Assessment, LLC. All rights reserved.

A one-way analysis of variance (ANOVA) showed a statistically significant difference in the mean difficulty of STAMP Spanish Reading items at the three major proficiency levels ($F [2,75] = 105.16$, $p < .001$). A Scheffé post-hoc test[4] additionally showed that the Advanced items ($M = 1.73$) were statistically significantly more difficult than the Intermediate items ($M = 0.01$), $p < 0.001$, Cohen's $d = 1.98$, which in turn were statistically significantly more difficult than the Novice items ($M = -1.82$), $p < 0.001$, Cohen's $d = 2.35$.

Cohen's d (Cohen, 1988) is an effect size and measure of the strength of the difference between two groups (in this case, items at two different proficiency levels). A Cohen's d difference of 2.35, as in the case between the average difficulty of intermediate and novice Spanish reading items, can be visualized below.



A Cohen's d of 2.35 can be interpreted in the following way, if we think of the items in the Novice and Intermediate levels of STAMP 4S Spanish Reading as being actual samples of items that Avant could insert into this test as a Novice or Intermediate item, respectively. There is a 99.06% chance (Cohen's U3) that any item randomly chosen from the Intermediate item pool for Spanish 4S Reading will be harder than the average difficulty of items in the Novice pool of items. Between the pools of items at the Novice and Intermediate levels, there is a 24% overlap in the distribution of their (IRT) difficulties. Finally, there is a 95.17% chance, if we take the pools of items at these two proficiency levels into consideration, that a randomly picked item at the Intermediate level will be harder than an item picked randomly from the Novice pool of items.

[4] A Scheffé post-hoc was employed due to the different number of items at each of the three ACTFL levels and because Scheffé is a more conservative test than Tukey HSD, therefore reducing the chance of detecting significant differences when in fact there are none.

Copyright © 2019 Avant Assessment, LLC. All rights reserved.

Figure 2 shows the average item difficulty for Novice, Intermediate, and Advanced items in the **Listening** section of STAMP 4S Spanish. A one-way analysis of variance (ANOVA) also detected a statistically significant difference in the mean difficulty of STAMP Spanish Listening items at the three major proficiency levels ($F [2,62] = 92.11$, $p < .001$). A Scheffé post-hoc test additionally showed that the Advanced items ($M = 1.38$) were statistically significantly more difficult than the Intermediate items ($M = - 0.01$), $p < 0.001$, Cohen's $d = 2.62$, which in turn were statistically significantly more difficult than the Novice items ($M = - 1.27$), $p < 0.001$, Cohen's $d = 1.93$.

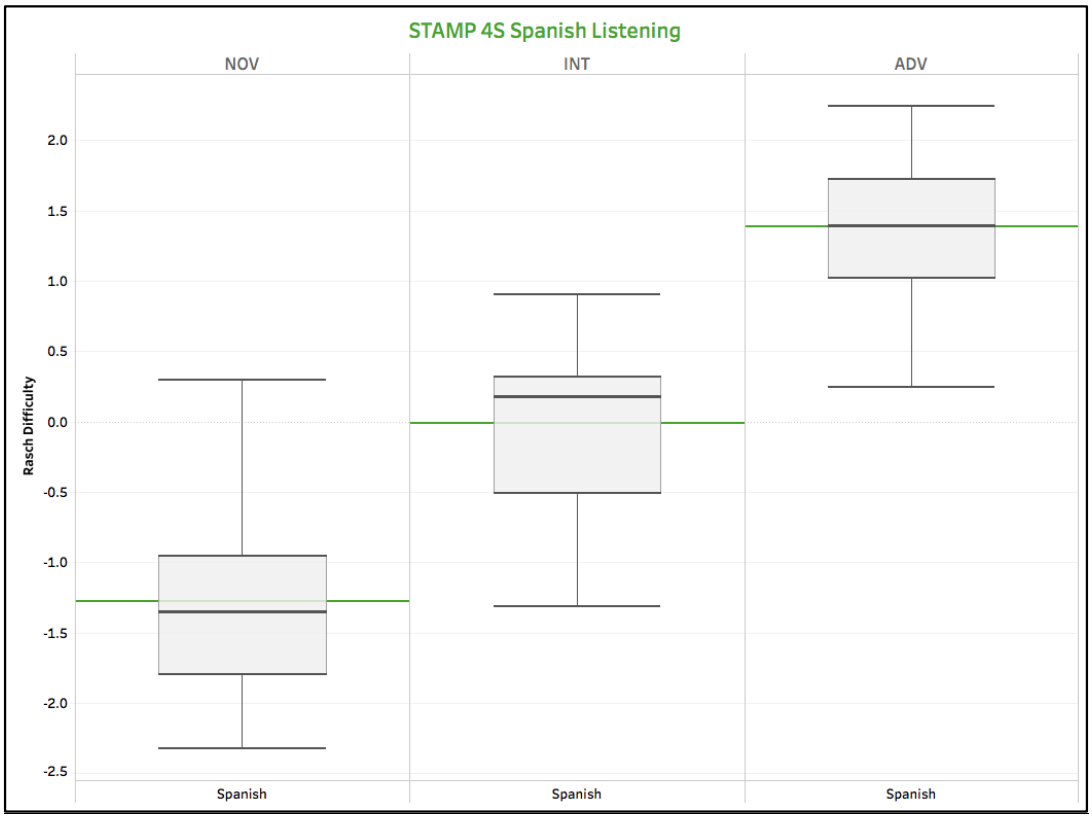


Figure 2. Mean item difficulty of Novice, Intermediate, and Advanced items in STAMP 4S Spanish (Listening).

The results above show clear and statistically significant differences between the mean difficulty of items written to target the Novice, Intermediate, and Advanced levels on the STAMP scale in the STAMP 4S Spanish, which provides support to the validity of the STAMP 4S Spanish test and the scores awarded to test-takers in these two sections.

In Reading, two of the Novice items are harder than one would expect and one of the Intermediate items is easier than expected. These three items are currently under review so that the appropriate course of action can be taken.

Copyright © 2019 Avant Assessment, LLC. All rights reserved.

French Reading and Listening

Figure 3 shows the mean item difficulty for Novice, Intermediate, and Advanced items in the **Reading** section of STAMP 4S French.

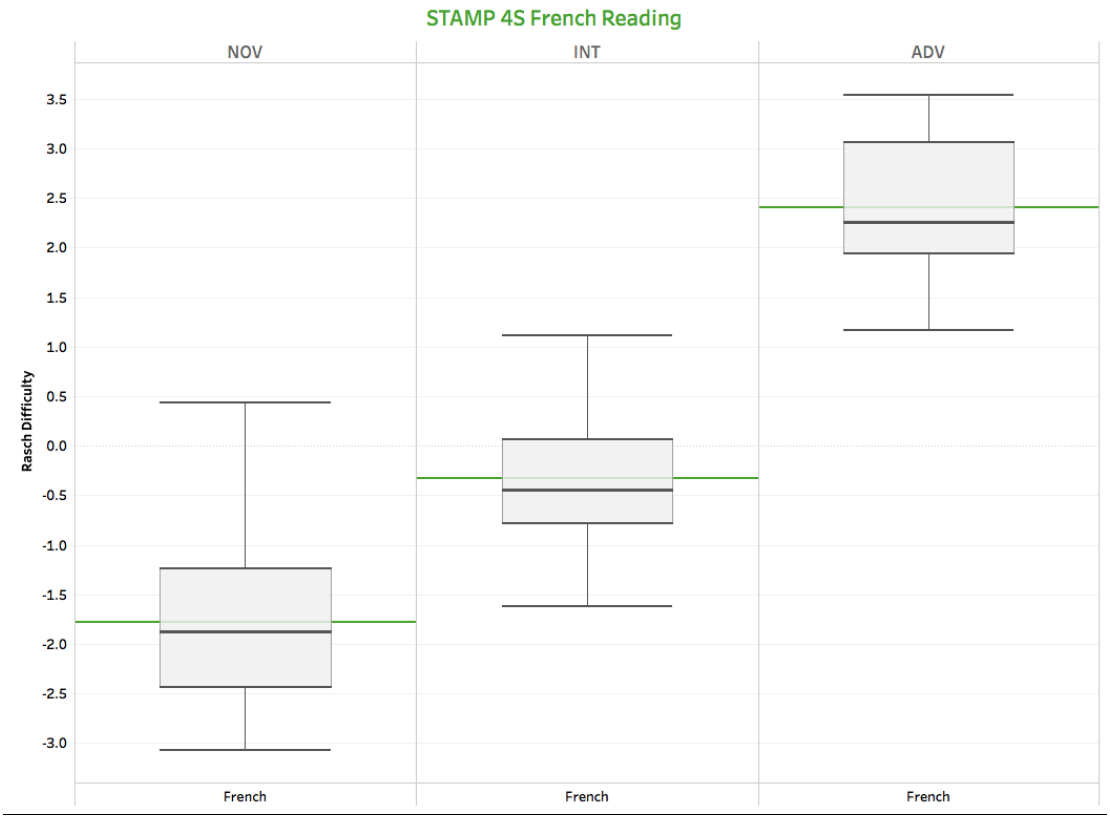


Figure 3. Mean item difficulty of Novice, Intermediate, and Advanced items in STAMP 4S French (Reading).

A one-way analysis of variance (ANOVA) showed a statistically significant difference in the mean difficulty of STAMP French Reading items at the three major proficiency levels ($F [2,61] = 169.65$, $p < .001$). A Scheffé post-hoc test additionally showed that the Advanced items ($M = 2.40$) were statistically significantly more difficult than the Intermediate items ($M = - 0.35$), $p < 0.001$, Cohen's $d = 4.16$, which in turn were statistically significantly more difficult than the Novice items ($M = -1.80$), $p < 0.001$, Cohen's $d = 1.91$.

Figure 4 shows the average item difficulty for Novice, Intermediate, and Advanced items in the **Listening** section of STAMP 4S French. A one-way analysis of variance (ANOVA) also detected a statistically significant difference in the mean difficulty of STAMP French **Listening** items at the three major proficiency levels ($F [2,53] = 60.85$, $p < .001$). A Scheffé post-hoc test additionally showed that the Advanced items ($M = 1.92$) were statistically significantly more difficult than the Intermediate items ($M = - 0.08$), $p < 0.001$, Cohen's $d = 3.39$ which in turn were statistically significantly more difficult than the Novice items ($M = - 1.05$), $p < 0.001$, Cohen's $d = 1.41$.

Copyright © 2019 Avant Assessment, LLC. All rights reserved.

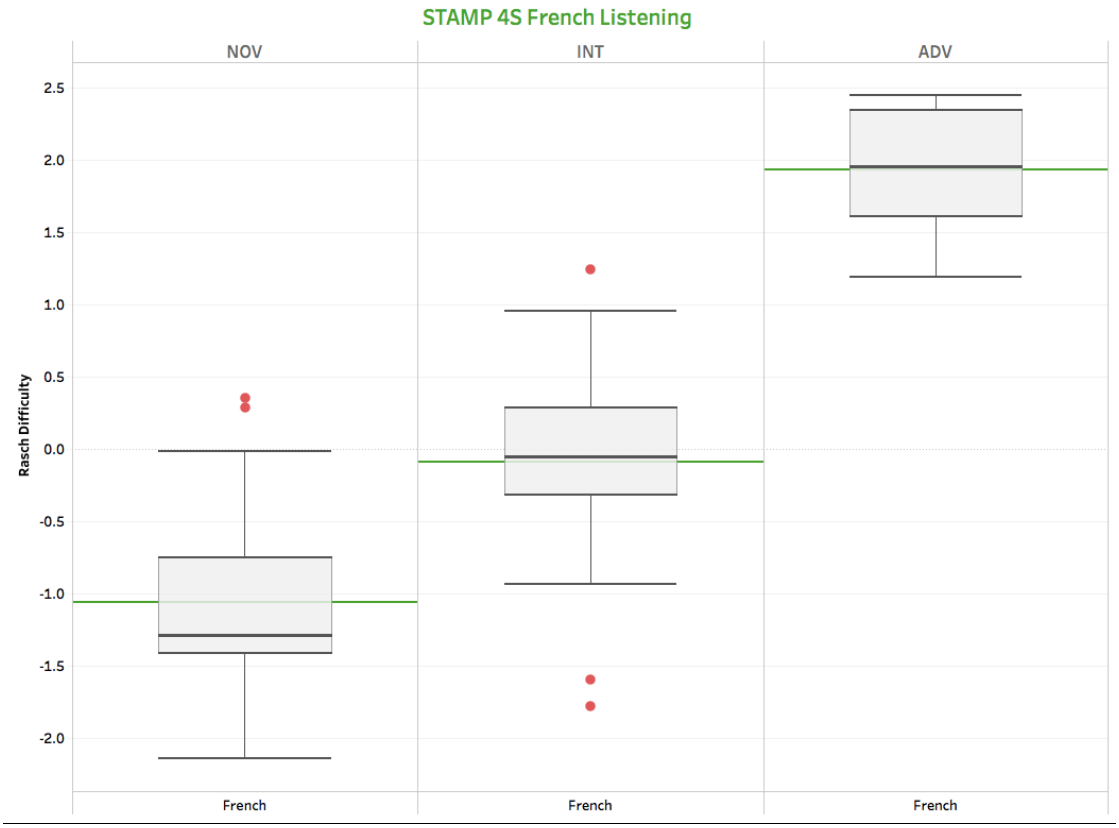


Figure 4. Mean item difficulty of Novice, Intermediate, and Advanced items in STAMP 4S French (Listening).

The results above show clear and statistically significant differences between the average difficulty of items written to target the Novice, Intermediate, and Advanced levels on the STAMP scale in STAMP 4S French, which provides support to the validity of the STAMP French test and the scores awarded to test-takers in these two sections.

In Listening, two of the Novice items seem considerably harder than one would expect. Conversely, two of the Intermediate items seem considerably easier than expected. These four items are currently under review.

Copyright © 2019 Avant Assessment, LLC. All rights reserved.

Japanese Reading and Listening

Figure 5 shows the average item difficulty for Novice, Intermediate, and Advanced items in the **Reading** section of STAMP 4S Japanese.

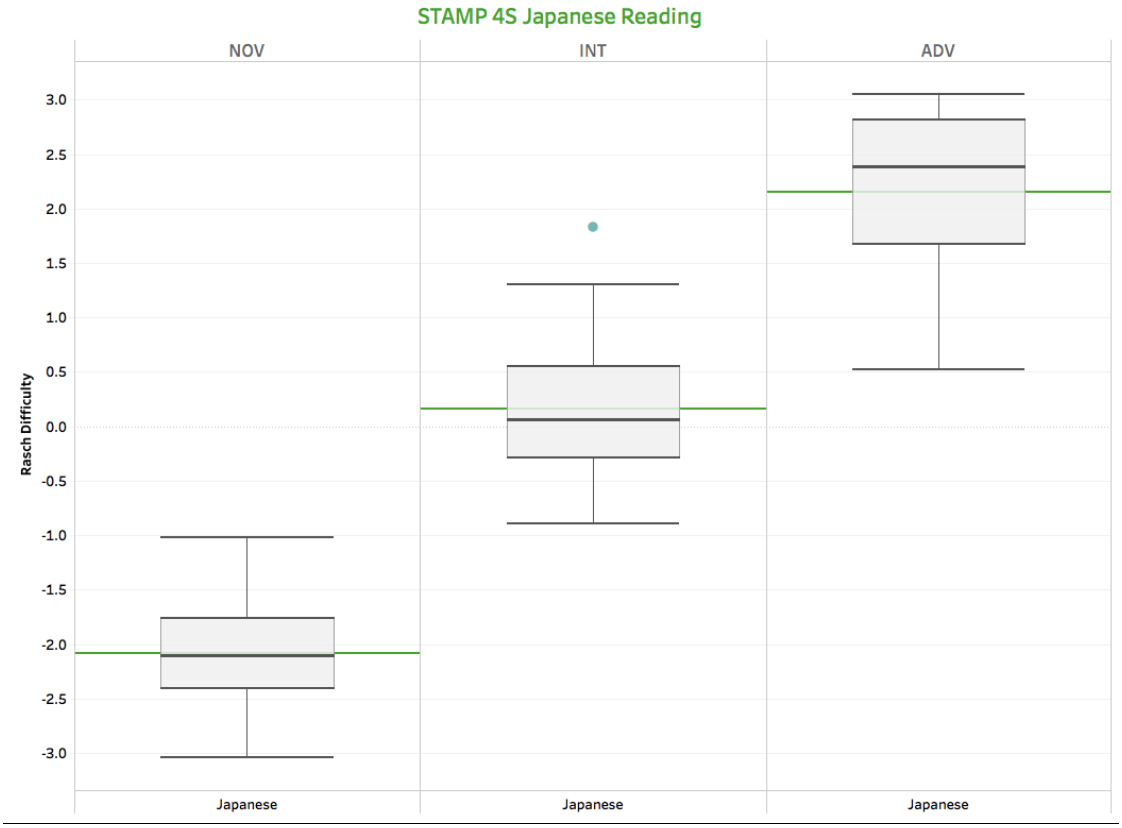


Figure 5. Mean item difficulty of Novice, Intermediate, and Advanced items in STAMP 4S Japanese (Reading).

A one-way analysis of variance (ANOVA) showed a statistically significant difference in the mean difficulty of STAMP Japanese Reading items at the three major proficiency levels ($F [2,59] = 193.94$, $p < .001$). A Scheffé post-hoc test additionally showed that the Advanced items ($M = 2.15$) were statistically significantly more difficult than the Intermediate items ($M = 0.16$), $p < 0.01$, Cohen's $d = 2.66$, which in turn were statistically significantly more difficult than the Novice items ($M = - 2.05$), $p < 0.001$, Cohen's $d = 3.42$.

Figure 6 shows the average item difficulty for Novice, Intermediate, and Advanced items in the **Listening** section of STAMP 4S Japanese. A one-way analysis of variance (ANOVA) also detected a statistically significant difference in the mean difficulty of STAMP Japanese Listening items at the three major proficiency levels ($F [2,59] = 71.49$, $p < .001$). A Scheffé post-hoc test additionally showed that the Advanced items ($M = 1.06$) were statistically significantly more difficult than the Intermediate items ($M = 0.06$), $p < 0.01$, Cohen's $d = 1.04$, which in turn were

Copyright © 2019 Avant Assessment, LLC. All rights reserved.

statistically significantly more difficult than the Novice items (M = - 2.00), $p < 0.001$, Cohen's d = 2.31.

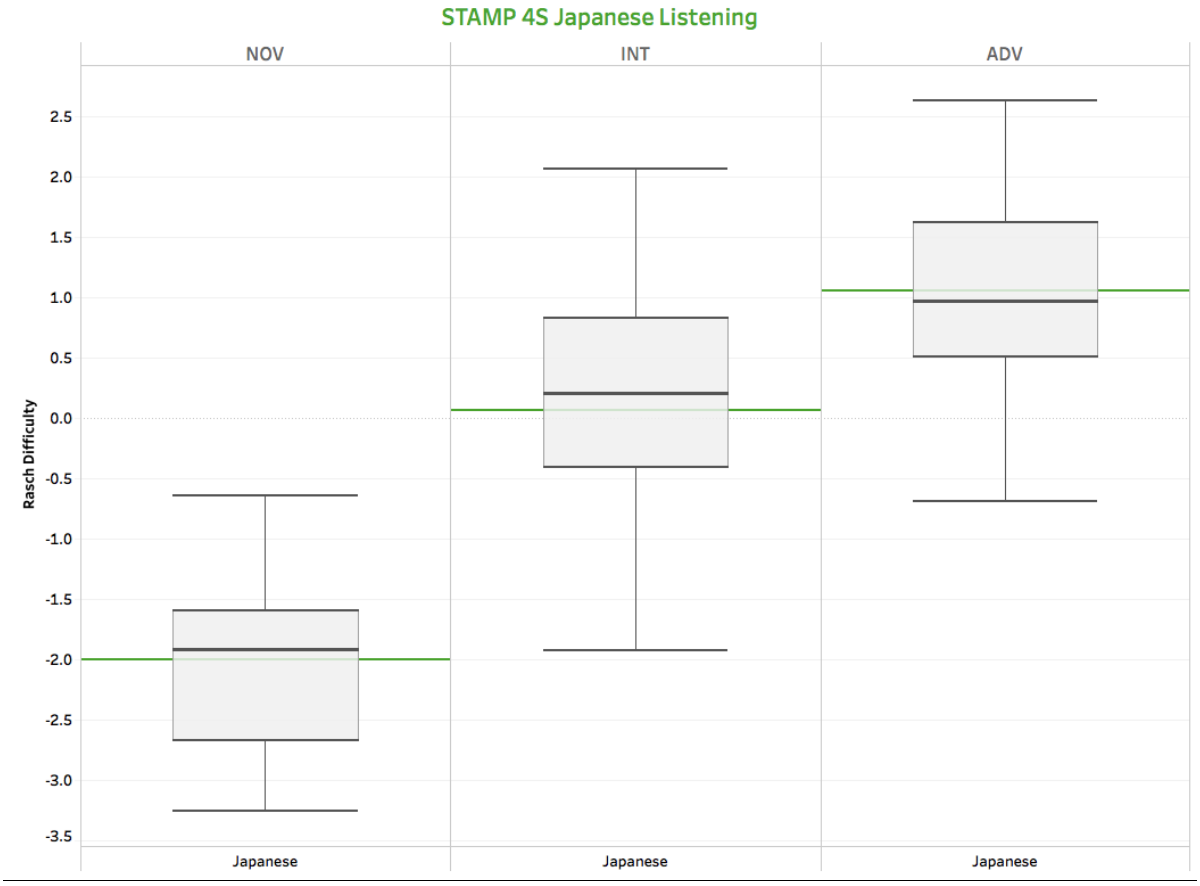


Figure 6. Mean item difficulty of Novice, Intermediate, and Advanced items in STAMP 4S Japanese (Listening).

The results above show clear and statistically significant differences between the average difficulty of items written to target the Novice, Intermediate, and Advanced levels on the STAMP scale in STAMP Japanese 4S, which provides support to the validity of the STAMP 4S Japanese test and the scores awarded to test-takers in these two sections.

In Reading, one Intermediate item is harder than one would expect. This item is currently under review at Avant, so that the appropriate course of action can be taken.

Copyright © 2019 Avant Assessment, LLC. All rights reserved.

Mandarin Chinese Reading (Simplified) and Listening

Figure 7 shows the average item difficulty for Novice, Intermediate, and Advanced items in the **Reading** section of STAMP 4S Mandarin Chinese Simplified.

Figure 7. Mean item difficulty of Novice, Intermediate, and Advanced items in STAMP 4S Mandarin Chinese Simplified (Reading).

A one-way analysis of variance (ANOVA) showed a statistically significant difference in the mean difficulty of STAMP Mandarin Chinese Simplified Mean items at the three major proficiency levels ($F [2,111] = 82.07$, $p < .001$). A Scheffé post-hoc test additionally showed that the Advanced items ($M = 1.51$) were statistically significantly more difficult than the Intermediate items ($M = 0.15$), $p < 0.001$, Cohen's $d = 1.47$, which in turn were statistically significantly more difficult than the Novice items ($M = -1.26$), $p < 0.001$, Cohen's $d = 1.33$.

Figure 8 shows the average item difficulty for Novice, Intermediate, and Advanced items in the **Listening** section of STAMP 4S Mandarin Chinese Simplified[5]. A one-way analysis of variance (ANOVA) also detected a statistically significant difference in the mean difficulty of STAMP 4S

[5] The Listening section in the STAMP 4S Mandarin Simplified test does not differ from the Listening section in the Mandarin Traditional test, given that the simplified vs traditional distinction only applies to writing.

Copyright © 2019 Avant Assessment, LLC. All rights reserved.

Mandarin Chinese Simplified Listening items at the three major proficiency levels (F [2,56] = 102.58, p < .001). A Scheffé post-hoc test additionally showed that the Advanced items (M = 2.44) were statistically significantly more difficult than the Intermediate items (M = 0.10), p < 0.001, Cohen's d = 2.06 which in turn were statistically significantly more difficult than the Novice items (M = - 1.88), p < 0.001, Cohen's d = 2.59.

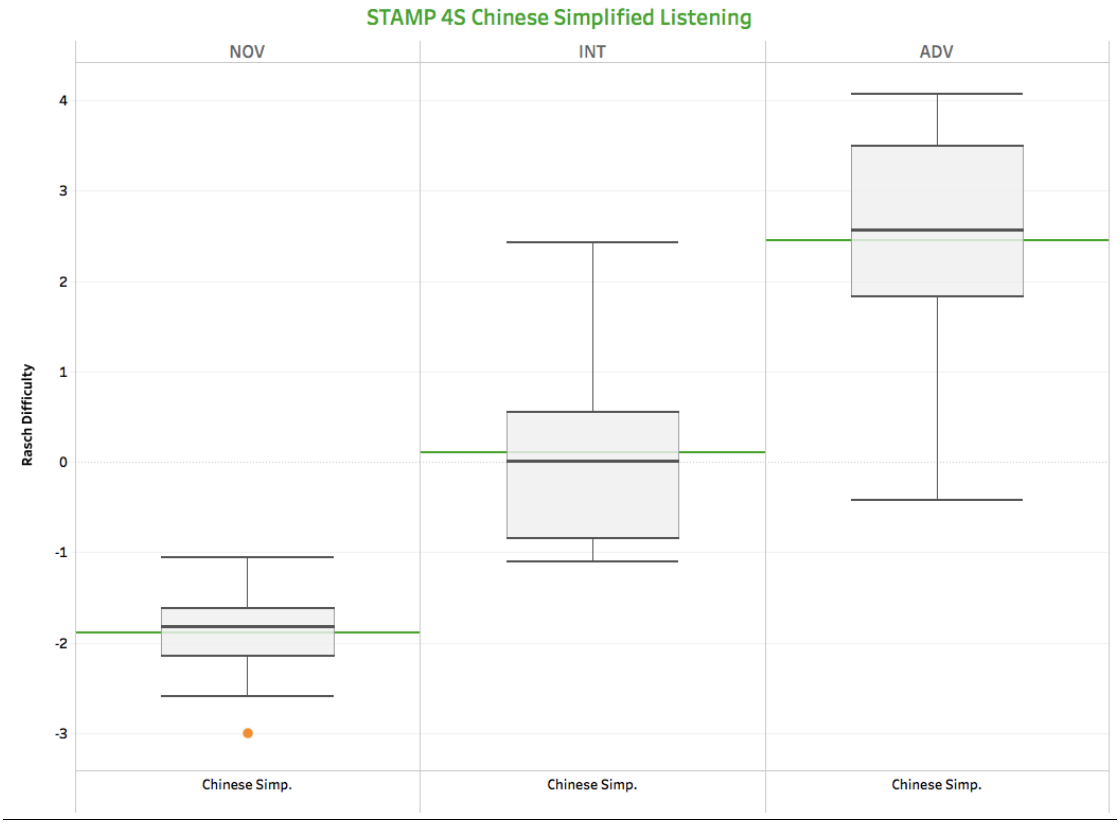


Figure 8. Mean item difficulty of Novice, Intermediate, and Advanced items in STAMP 4S Mandarin Chinese Simplified (Listening).

The results above show clear and statistically significant differences between the average difficulty of items written to target the Novice, Intermediate, and Advanced levels on the STAMP scale in STAMP 4S Mandarin Chinese Simplified, which provides support to the validity of the STAMP 4S Mandarin Chinese Simplified test and the scores awarded to test-takers in these two sections.

In Reading, one Novice item is much easier than others at the same level. Although this is not an issue and there is value in having significantly easier Novice items in the test (better targeted to students at the Novice-Low level), we are currently examining this item in order to better understand the reasons why it is so easy for test-takers. The same applies to the one outlier item at the Novice level in Listening.

Copyright © 2019 Avant Assessment, LLC. All rights reserved.

Arabic Reading and Listening

Figure 9 shows the average item difficulty for Novice, Intermediate, and Advanced items in the **Reading** section of STAMP 4S Arabic.

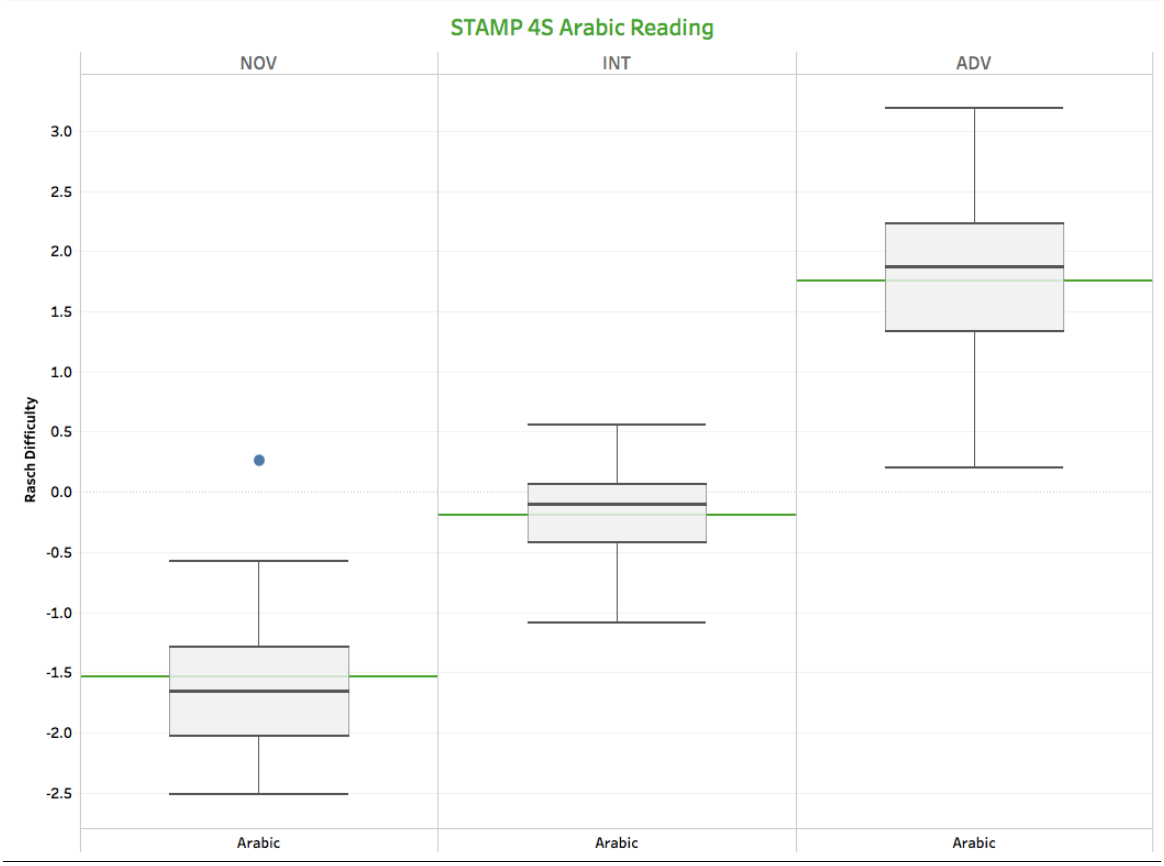


Figure 9. Mean item difficulty of Novice, Intermediate, and Advanced items in STAMP 4S Arabic (Reading).

A one-way analysis of variance (ANOVA) showed a statistically significant difference in the mean difficulty of STAMP 4S Arabic Reading items at the three major proficiency levels ($F [2,54] = 117.50$, $p < .001$). A Scheffé post-hoc test additionally showed that the Advanced items ($M = 1.75$) were statistically significantly more difficult than the Intermediate items ($M = - 0.19$), $p < 0.001$, Cohen's $d = 2.99$, which in turn were statistically significantly more difficult than the Novice items ($M = -1.53$), $p < 0.001$, Cohen's $d = 2.28$.

Figure 10 shows the average item difficulty for Novice, Intermediate, and Advanced items in the **Listening** section of STAMP 4S Arabic. A one-way analysis of variance (ANOVA) also detected a statistically significant difference in the mean difficulty of STAMP Arabic Listening items at the three major proficiency levels ($F [2,65] = 104.59$, $p < .001$). A Scheffé post-hoc test additionally showed that the Advanced items ($M = 1.81$) were statistically significantly more difficult than the

Copyright © 2019 Avant Assessment, LLC. All rights reserved.

Intermediate items (M = 0.04), $p < 0.001$, Cohen's d = 2.43, which in turn were statistically significantly more difficult than the Novice items (M = - 1.61), $p < 0.001$, Cohen's d = 1.86.

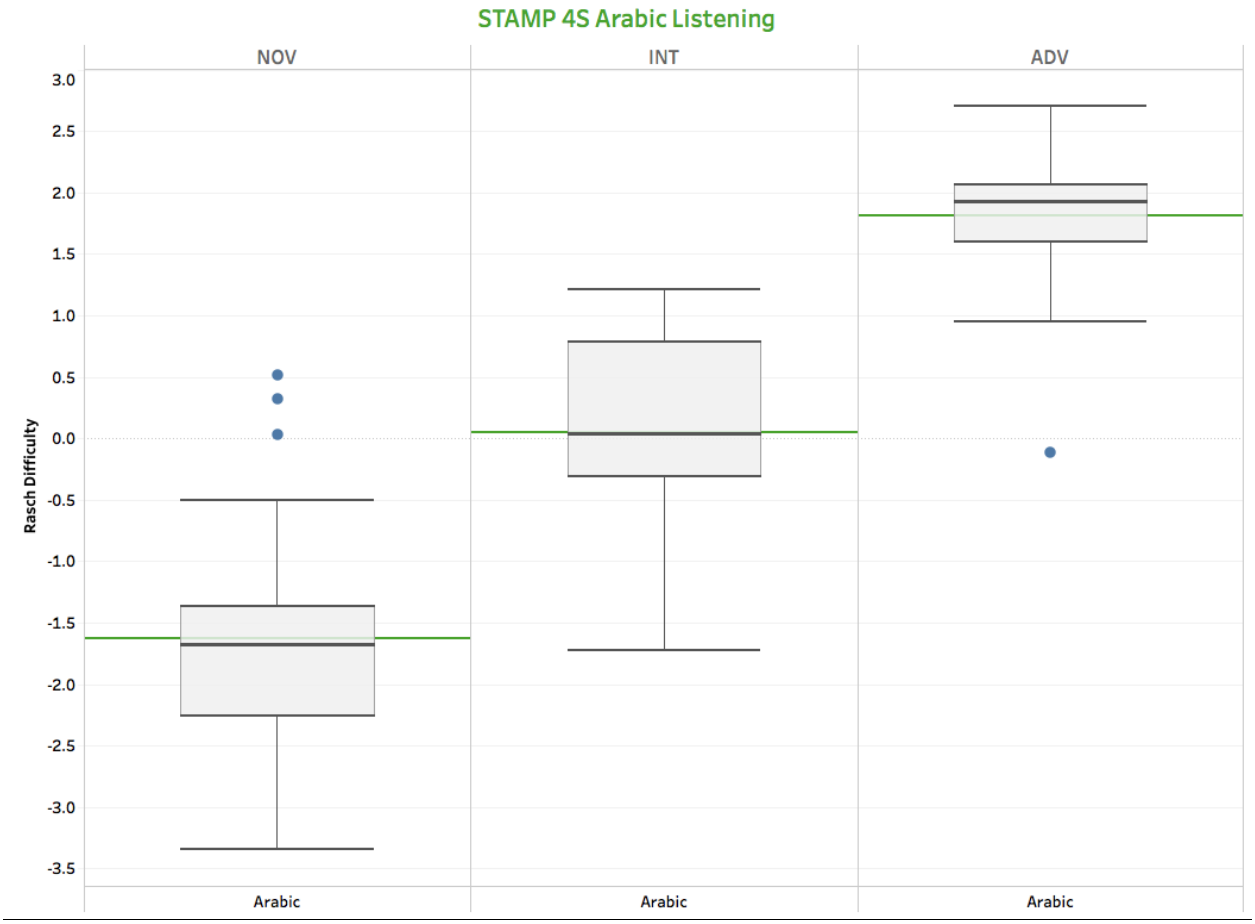


Figure 10. Mean item difficulty of Novice, Intermediate, and Advanced items in STAMP 4S Arabic (Listening).

The results above show clear and statistically significant differences between the average difficulty of items written to target the Novice, Intermediate, and Advanced levels on the STAMP scale in STAMP 4S Arabic, which provides support to the validity of the STAMP 4S Arabic test and the scores awarded to test-takers in these two sections.

In Reading, one Novice item is considerably harder than others at the same level. This item is currently under review. In Listening, three Novice items are harder than one would expect for this level and one Advanced item is easier than one would expect. These four items are also currently under review.

Before proceeding to answering Question #2, we can find in Figure 11 a visual summary with the mean difficulty of Novice, Intermediate, and Advanced items in Reading and Listening for

Copyright © 2019 Avant Assessment, LLC. All rights reserved.

the five languages above. The information is the same already presented, but this different visualization may be helpful for some readers.

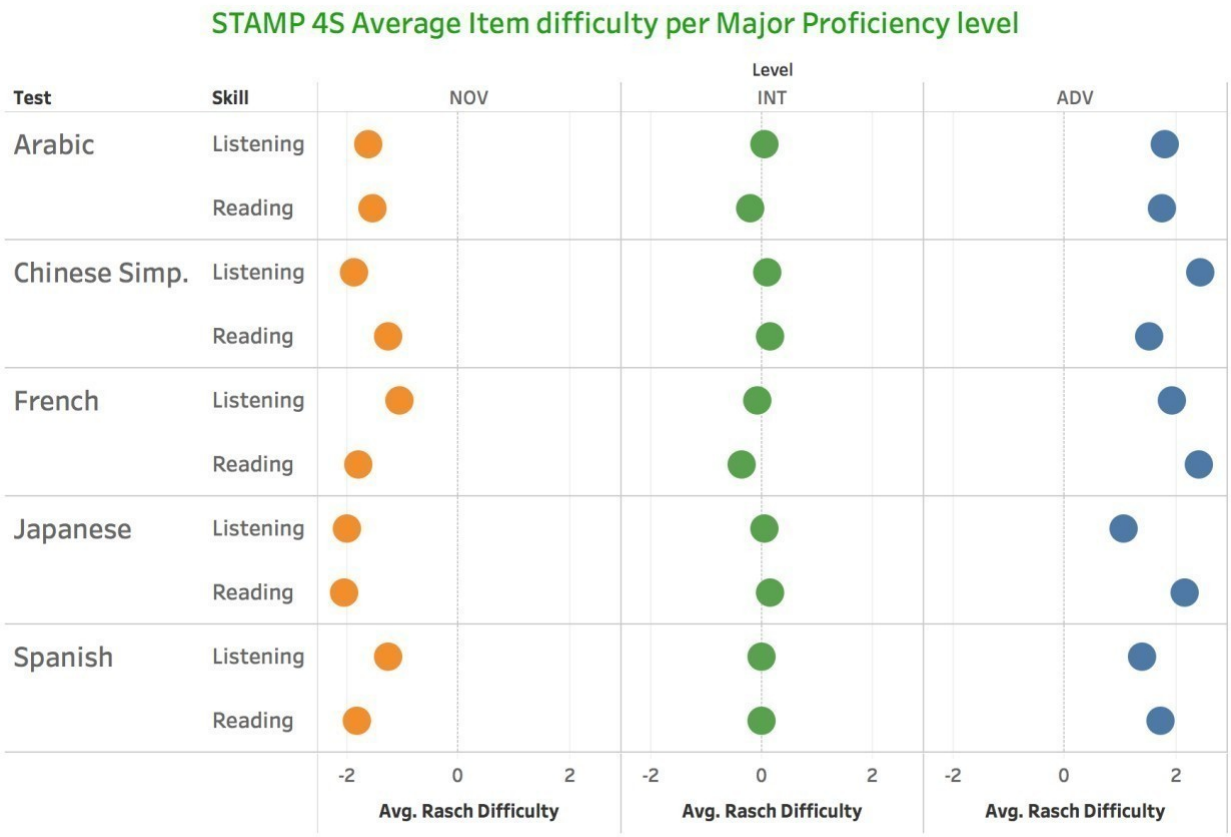


Figure 11. Mean item difficulty of Novice, Intermediate, and Advanced items in five STAMP 4S languages across Reading and Listening (each dot signifies the average item difficulty of all items at each level and domain for the represented languages)

We can see from Figure 11 that the mean difficulty of items at the Novice, Intermediate, and Advanced levels in either Reading or Listening never cross the space of the mean difficulty of items at an adjacent level in any other language or test skill. Although each language/skill must be calibrated on a different IRT scale and different languages will always behave differently due to intrinsic characteristics of the language itself as well as due to the specific test-taking population, a Rasch value of 0 still holds the same meaning across all scales: it represents the mean difficulty for all items in the pool for that specific language/skill (which will include items at all three major proficiency levels). Therefore, the fact that the average of items at one level for a given language/skill never crosses into the space observed for the average of an adjacent skill provides support that items at the three major proficiency levels behave similarly enough across different languages and skills, since they are written based on the same proficiency scale (the STAMP scale, which is based on the ACTFL Proficiency Guidelines).

Copyright © 2019 Avant Assessment, LLC. All rights reserved.

Question 2: Do test-takers who receive a higher STAMP score indeed have a higher level of proficiency in the language?

Another important piece of evidence in support of the validity of a test built on the ACTFL proficiency scale such as the STAMP 4S test is whether test-takers who score at higher ACTFL levels indeed have higher levels of language proficiency, as demonstrated by empirical results. Applying a certain proficiency label to a test score is very simple to do. The hard part involves being able to show that the proficiency level awarded at the end of a test administration is meaningful[6] and that there is also an empirical hierarchy of actual language proficiency between test-takers at the various proficiency levels awarded by the test. It is to be expected that test-takers at higher levels on the STAMP scale (e.g., Advanced Low) have higher empirical language proficiency than test-takers at relatively lower levels on the STAMP scale (e,g., Intermediate Mid). This hierarchy of ability must be found across the entire spectrum of scores on the test. In the case of the STAMP 4S, the spectrum of language proficiency goes from level 1 (Novice Low) to Level 9 (Advanced High).

Using results data from the very same test-takers employed to answer Question #1 above, let's examine if the hypothesis of higher empirical proficiency for higher STAMP levels holds when we look at actual assessment results. For this analysis, we are employing results from STAMP 4S Reading and Listening as an example, since this is the test with the highest number of test-takers for any given year.

[6] This includes, among others, being able to show that the proficiency of students who score at level X on STAMP corresponds to descriptors for level X on the ACTFL's Proficiency Guidelines.

Copyright © 2019 Avant Assessment, LLC. All rights reserved.

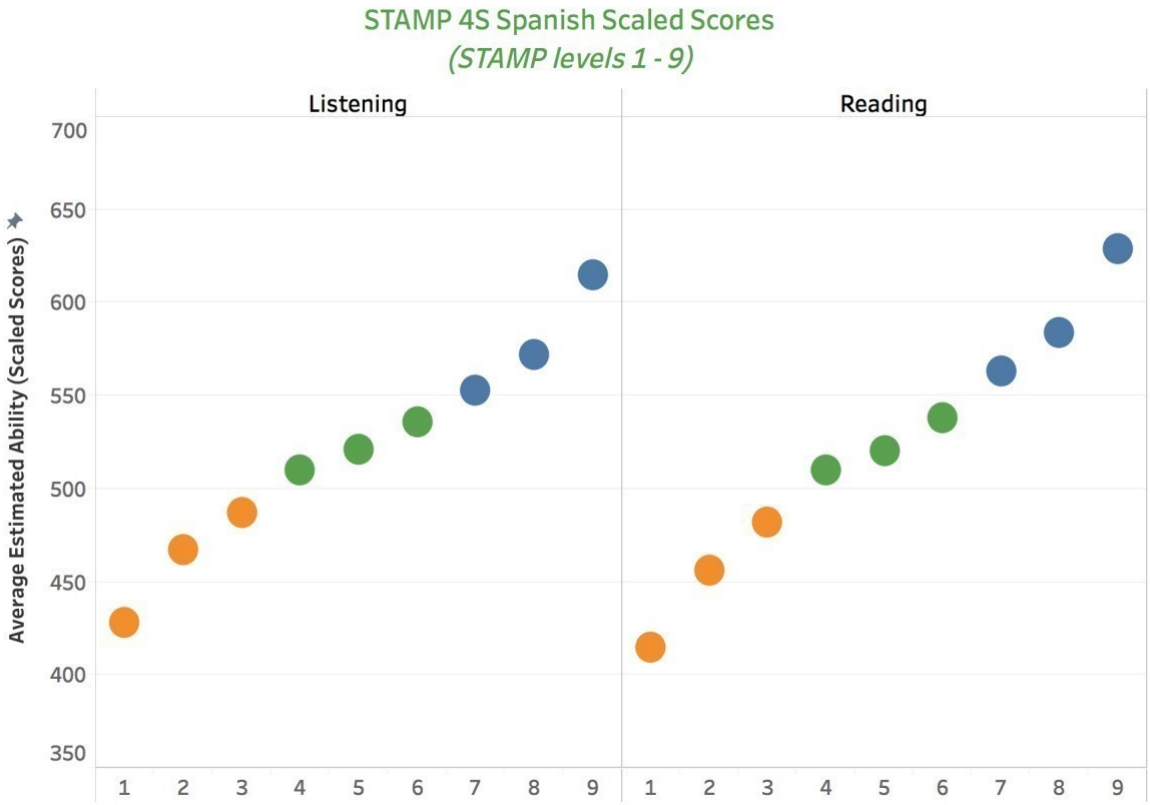


Figure 12. Average empirical language ability of test-takers from STAMP Level 1 (Novice-Low, bottom orange circle) to STAMP Level 9 (Advanced High, top blue circle) for STAMP 4S Spanish Reading and Listening.

We can see from Figure 12 that the average empirical proficiency (represented by scaled scores[7]) of test-takers at each of the nine STAMP levels monotonically increases, from level 1 all the way to level 9. This finding provides support to the validity-related claim that STAMP 4S scores are meaningful and that STAMP scores can indeed be used to compare the proficiency level of one student in relation to that of another. In other words, a higher STAMP level means higher proficiency in the language.

It is important to keep in mind that scaled scores are section- and language-dependent. Scaled scores for Spanish Reading, for example, cannot be compared directly with scaled scores in Spanish Listening or French Reading, since these are on different scales, which is a requirement of IRT. Therefore, a scaled score of 550 in one section/language can (and often does) indicate a different level of proficiency than the exact same scaled score in a different section/language combination.

[7] The scaled scores in the STAMP test are simply a linear scaling (transformation) of the IRT ability the system assigns to each test-taker at the end of the test, based on the items they encountered during the test, the difficulty of each of those items, and the test-taker's response to each of those items.

Copyright © 2019 Avant Assessment, LLC. All rights reserved.

DISCUSSION

In assessing the validity of an assessment that reports scores to test-takers on the STAMP scale, which is based on the ACTFL Proficiency Guidelines, it is crucial that the following two characteristics of the test be supported through analysis of actual test data: (a) items written to target the Advanced level are harder, on average, than items written to target the Intermediate level, which in turn are harder than items written to target the Novice level, and (b) test-takers who receive higher scores on the STAMP scale indeed have higher language proficiency than test-takers who receive scores that are lower on the STAMP scale. In the case of the STAMP 4S tests, we have shown evidence that clearly warrants both of these assumptions.

We believe the reasons why these assumptions are supported in the case of STAMP 4S are multifold, but perhaps the most important reason lies in Avant's approach to test development and ongoing quality assurance of its item pool. The behavior of every item in a STAMP test is closely monitored by trained Avant personnel (which includes measurement specialists and item writers) from the moment it is written, pilot tested, and made operational in a STAMP test, all the way to when the item may be retired so as to allow the introduction of new items into the test. At Avant, content-related decisions are triangulated with statistical-based decisions, so that the majority of items written to target a certain STAMP level will tend to also behave accordingly statistically. If that is not the case, the item will certainly be detected during STAMP's annual refreshes and the appropriate course of action will be taken in order to continue to ensure the high quality and validity of the STAMP assessment.

REFERENCES

ACTFL. (2012). ACTFL Proficiency Guidelines. Retrieved October 10, 2019, from ACTFL: https://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Cox, T. L.,&Clifford, R. (2014). Empirical validation of listening proficiency guidelines. Foreign Language Annals, 47, 379–403.

Hendrickson, A. 2007. An NCME instructional module on multistage testing. Educational Measurement: Issues and Practice 26: 44–52.

Linacre, J. M. (2000). Computer-adaptive testing: A methodology whose time has come. In S. Chae, U. Kang, E. Jeon, & J. M. Linacre (Eds.): Development of Computerised Middle School Achievement Tests. MESA Research Memorandum, 69. Seoul, South Korea: Komesa Press.

Meyer, P. (2014). Applied measurement with Jmetrik. New York, NY: Routledge.

Copyright © 2019 Avant Assessment, LLC. All rights reserved.

Thompson, N. A. (2011). Advantages of Computerized Adaptive Testing [White paper]. Retrieved october 10, 2019, from Assessment Systems: https://assess.com/docs/Advantages-of-CAT-Testing.pdf

Wainer, H., Kaplan, B., & Lewis, C. (1992). A comparison of the performance of simulated hierarchical and linear testlets. Journal of Educational Measurement, 29, 243–251.

Weiss, D. J. (1985). Adaptive testing by computer. Journal of Consulting and Clinical Psychology, 53(6), 774.

Yan, D., C. Lewis and M. Stocking. 2004. Adaptive testing with regression trees in the presence of multidimensionality. Journal of Educational and Behavior Statistics 29: 293–316.

Copyright © 2019 Avant Assessment, LLC. All rights reserved.

APPENDIX

Descriptive statistics[8] for items at all three major ACTFL levels for Reading and Listening across Spanish, French, Japanese, Mandarin Chinese (Simplified), and Arabic

Spanish (Reading, Listening)				
Intended Level	**N**	**M**	**95% CI** (lower to upper)	**SD**
Novice	21 22	- 1.82 - 1.27	- 2.15 to - 1.49 - 1.60 to - 0.94	0.72 0.74
Intermediate	32 25	0.01 - 0.01	-0.29 , 0.32 - 0.24 to 0.22	0.83 0.56
Advanced	25 18	1.73 1.38	1.36 to 2.10 1.13 to 1.63	0.90 0.49

French (Reading, Listening)				
Intended Level	**N**	**M**	**95% CI** (lower to upper)	**SD**
Novice	24 23	- 1.80 - 1.05	- 2.16 to - 1.38 - 1.34 to - 0.76	0.84 0.67
Intermediate	22 25	- 0.35 - 0.08	-0.65 , 0 - 0.37 to 0.20	0.66 0.70
Advanced	18 8	2.40 1.92	2.05 to 2.76 1.55 to 2.30	0.66 0.45

[8] All M, 95% CI, and SD values are rounded down to two decimals.

Copyright © 2019 Avant Assessment, LLC. All rights reserved.

Japanese (Reading, Listening)				
Intended Level	**N**	**M**	**95% CI** (lower to upper)	**SD**
Novice	23 23	- 2.05 - 2.00	- 2.34 to - 1.82 - 2.36 to - 1.62	0.56 0.71
Intermediate	21 18	0.16 0.06	- 0.16 to 0.49 - 0.45 to 0.58	0.72 1.04
Advanced	18 21	2.15 1.06	1.76 to 2.53 0.64 to 1.47	0.77 0.86

Mandarin Chinese (Simpl.) (Reading, Listening)				
Intended Level	**N**	**M**	**95% CI** (lower to upper)	**SD**
Novice	47 23	- 1.26 - 1.88	- 1.57 to - 0.94 - 2.08 to - 1.69	1.07 0.45
Intermediate	31 21	0.15 0.10	- 0.23 to 0.53 - 0.36 to 0.58	1.04 0.98
Advanced	36 15	1.51 2.44	1.25 to 1.78 1.74 to 3.15	0.78 1.27

Arabic (Reading, Listening)				
Intended Level	**N**	**M**	**95% CI** (lower to upper)	**SD**
Novice	23 26	- 1.53 - 1.61	- 1.85 to - 1.20 - 2.02 to - 1.22	0.71 0.95
Intermediate	15 20	- 0.19 0.04	- 0.42 to 0.04 - 0.33 to 0.43	0.43 0.81
Advanced	19 22	1.75 1.81	1.36 to 2.15 1.53 to 2.09	0.81 0.63

Copyright © 2019 Avant Assessment, LLC. All rights reserved.

INFO
☎ (888) 731-7887
info@avantassessment.com

OFFICE
☎ (541) 338-9090
940 Willamette Street, Suite 530
Eugene, OR 97401 USA
Map & Directions

SUPPORT
5:00 am - 5:00 pm Pacific Time M-F
☎ (888) 713-7887
support@avantassessment.com