Reliability and Accuracy of Ratings in the Writing and Speaking sections of STAMP tests

Victor D. O. Santos, *Ph.D*.
Director of Assessment and Research
Avant Assessment

September 2022

ABSTRACT

The STAMP 4S and STAMP WS tests within the STAMP (Standards-Based Measurement of Proficiency) family of assessments include a Writing and a Speaking section. A crucial piece of evidence for the validity of the scores in these sections, given their intended uses and interpretations, comes from the extent to which the scores can be shown to be reliable and accurate. In this paper, we show the results of a recent analysis conducted on the ratings in the Writing and Speaking sections across five representative STAMP 4S languages (Arabic, Spanish, French, Chinese Simplified, and Russian) and three representative STAMP WS languages (Amharic, Haitian Creole, and Vietnamese). The results, based on the analysis of over 23,000 examinee responses across these eight languages, show a high level of scoring accuracy and reliability for both the Writing and Speaking sections of STAMP, thus providing strong support for the validity of the scores from these sections given their intended interpretations and uses.

The Writing and Speaking Sections of STAMP

The STAMP family of tests (Standards-Based Measurement of Language Proficiency) assess real-world language proficiency and are aligned to the ACTFL proficiency guidelines. The STAMP 4S test is a four-skill test of language proficiency, accredited by the American Council on Education (ACE), and available at the time of this writing in 14 languages. The STAMP WS, also accredited by ACE, is a test of language proficiency in the two productive skills of Writing and Speaking, and is available at the time of this writing in 24 languages.

Two important factors in assessing the extent to which scores from a test can be said to be valid, given what a test purports to measure and the intended uses of those test scores, are the reliability and accuracy of the test scores. In this short paper, we will discuss and examine the reliability and accuracy of ratings for the Writing and Speaking sections of STAMP, in which trained human raters must assign a STAMP level between 0 (No proficiency) and 8 (Advanced-Mid) to examinees' language productions.

In each of the Writing and Speaking sections of the STAMP test, examinees are given three real-world scenario-based prompts to which they must respond. Examinees are instructed to write as much as possible and to “show off” their language skills to the best of their ability. An examinee’s response to each of the three prompts in the section is scored by Avant-certified raters, who must pass a thorough and stringent training and certification program in order to be allowed to rate STAMP responses. After these raters start rating real, operational STAMP Writing and Speaking responses, Avant and its rater managers keep a close eye on how each rater is performing by means of qualitative and quantitative measures to ensure the high-quality of our ratings and ensure that each and all Avant raters are rating to the company’s established standards.

In 80% of the time, a Speaking or Writing response is rated by a single Avant rater. The score/STAMP level assigned to that response by the rater becomes the official score for that response in the system. In 20% of the time, a response is rated by at least two Avant raters. When the two raters agree on the STAMP score, that becomes the official score assigned to that response. In case the two raters disagree, an Avant rater manager is brought in to rate the response. The rating provided by the rater manager becomes the official score assigned to that response. The rating of each response is done completely independently of the examinee’s response to the other two responses. When rating a given response, raters do not have access to any information regarding the examinee, their score on their other responses for that skill, or to the score awarded to that response by any other rater, all of which increase the validity of the rating for each response.

An examinee’s final STAMP score for either the Writing or the Speaking section is calculated based on the specific STAMP level they received for each of the three prompts they responded to. The official STAMP level awarded for the section becomes the highest proficiency level the examinee was able to sustain (*i.e.*, demonstrate in at least two instances) across their three responses.

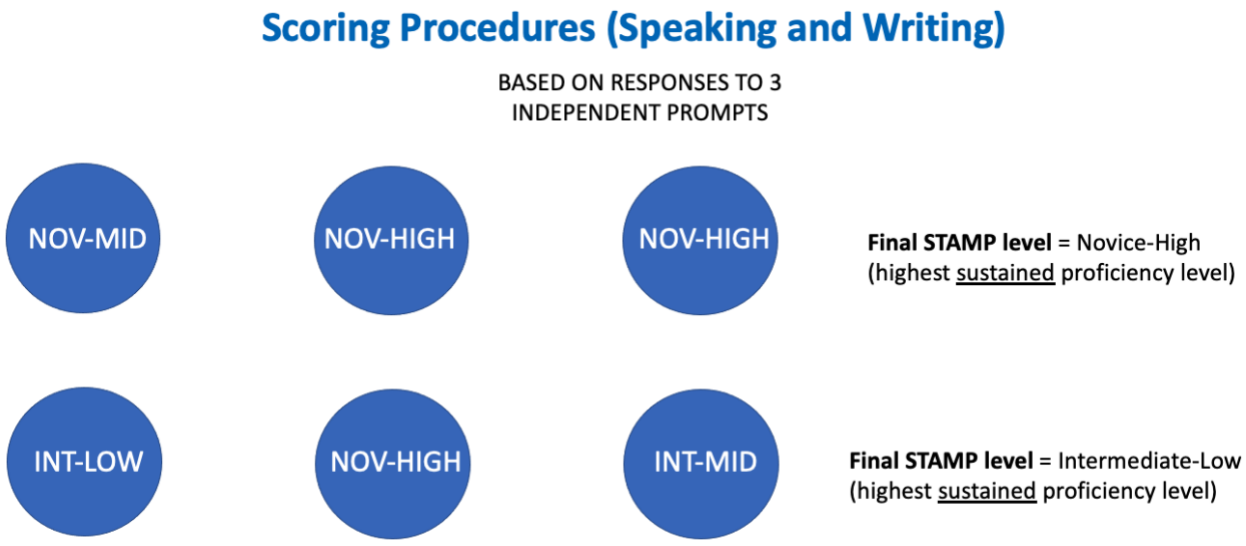

Figure 1. System rules for arriving at an examinee's final STAMP level for the Writing and Speaking sections

As shown in Figure 1, if an examinee is awarded a Novice-Mid for their first response, a Novice-High for their second, and a Novice-High for their third, that examinee's official STAMP level for that section becomes a STAMP 3 (Novice-High) since that is the highest level of proficiency they were able to sustain in at least two instances. Alternatively, if they are awarded an Intermediate-Low for their first response, a Novice-High for their second, and an Intermediate-Mid for their third, their final STAMP level in that section becomes an Intermediate-Low, which is the highest level they were able to sustain in at least two instances (the first and third, in this case).

The use of three independent prompts in the Writing section and three independent prompts in the Speaking section of STAMP has two main advantages. The first advantage is that it allows examinees to be assessed on different topics, thus supporting the premise that the level of proficiency awarded at the end of the section will generalize to other scenarios in the real-world. The second advantage is that, coupled with the scoring methodology described above, it helps to minimize the effect of any possible rating bias by any individual Avant rater.

We now turn our attention to the definition of reliability and accuracy.

Reliability

Reliability can be defined as "consistency of measurement" (Bachman & Palmer, 1996). Simply put, it is the extent to which scores on a given test can be trusted (*relied* upon) to stay the same

if an examinee were to take that test again over different occasions or take different forms of the test, assuming that the examinee's proficiency in what the test measures has not changed in the meantime.

For instance, if an examinee takes a language proficiency test today and receives a score of Intermediate-Low but then receives a score of Intermediate-High on the same test tomorrow, we could assume, provided that the examinee's knowledge of the language and their mental state has not changed, that the test may not be highly reliable. Along the same lines, if an organization makes a test available in various parallel forms (usually done to increase test security) but an examinee receives a score of Advanced-Low on one form and then Intermediate-Mid on another form, we can once again assume there may be a lack of consistency in measurement, and therefore an issue of lack of reliability, with that test.

One of the factors that contribute to the reliability of a test is the manner in which the test is scored. In the STAMP test, the Reading and Listening sections are composed of multiple-choice questions and the examinee's responses are automatically scored by a computerized system. What this means is that if an examinee provides the same response to the same items on different occasions, they will always receive the same score.

On the other hand, the Writing and Speaking sections of STAMP are scored by human raters. Therefore, it is feasible that an examinee could receive a different score, for the same exact response, depending on who happens to be rating that examinee's response. Of course, the more well trained raters are, the less we would expect scores to vary due to differences in leniency, strictness, or any possible bias on the part of raters.

Accuracy

Examinees expect their score on a test to only be dependent on how much or how little they have of the construct being measured by the test (in the case of STAMP, proficiency across each of the language domains). Accuracy relates to the extent to which the score awarded to an examinee's response correctly describes their ability in that construct. As such, if an examinee submits a Speaking response at the Intermediate-High level but the two raters who assigned a level to that response assign an Intermediate-Low, we could say that this is an inaccurate score. If the two other raters were to rate that same response two months later and assigned it Intermediate-Low as well, the scores would once again be inaccurate, despite being reliable (not having changed from one occasion to the next or from one rater to the next).

Figure 2 describes the difference between reliability and accuracy. Naturally, we would like tests to be both reliable and accurate. Having these two conditions met provides strong support for the validity of the test scores and their intended uses.

Figure 2: Reliability and Accuracy (source: Matrix Education)

Statistics Commonly Used to Evaluate the Reliability and Accuracy of Scores by Raters

When examinee responses on a test are scored by human raters, as in the case of STAMP, it is important to ascertain that the scores reflect the quality of the response itself and are therefore not impacted (or only minimally impacted) by the profile of the specific rater (or raters) who happens to evaluate that response. In other words, the score should be dependent only on how much of the construct measured by the test a certain examinee may demonstrate in their response and not on how lenient, strict, or biased a rater may be.

Statistics are often offered by language test providers to show the extent to which the scores awarded by human raters to examinees' responses may be affected by who happens to be doing the rating. Often in the language testing literature, these statistics are provided by comparing the ratings that two separate raters would give to the same essay. It is assumed that it is highly desirable that any two raters should assign the same score as often as possible to the same essay, which would show that the rating process is a highly reliable one.

However, as we have seen above, reliability must be accompanied by accuracy and the latter should also be investigated. After all, two random raters may assign the same score to an essay but both could be wrong. In a well-developed and well-scored test, the ideal scenario is when raters highly agree with one another *and* happen to be correct (accurate) in the scores they assign to responses.

It is important to understand that it is not viable to always expect perfect agreement between two human raters. Despite all the training they each may have gone through and all the experience and expertise each one may have regarding the construct being evaluated (in our case, language proficiency), even highly qualified humans disagree at times. Doctors do it. Engineers do it. Scientists do it. Therefore, the idea is to aim for as high an agreement as is feasible, and which proves defensible given the uses and interpretations of the scores from that test.

Below are the statistical measures that we at Avant Assessment run on the STAMP test in order to assess the quality of the rating provided by our team of human raters. While many companies may only report exact and adjacent agreement, we assess our raters on additional measures as well, since any specific measure can only provide partial information as to the quality of the raters. The more measures included, the more we are able to triangulate the results and arrive at a conclusive decision. The measures we will report in this paper are:

Exact Agreement:

This measure is reported as a percentage that indicates the percentage of times, across the entire dataset analyzed, when the level awarded to a given response by Rater 1 is exactly the same as the level awarded by Rater 2. For example, if Rater 1 awards a STAMP level 5 to a response and Rater 2 also awards a STAMP level 5 to that same response, that would be considered an instance of exact agreement. Feldt and Brennan (1989) suggest that when two raters are used, there should be an exact agreement of at least 80%, with 70% being considered acceptable for operational use.

Exact + Adjacent Agreement:

This measure is reported as a percentage that indicates the percentage of times, across the entire dataset analyzed, when the level awarded to a given response by Rater 1 is either exact or adjacent to the level awarded by Rater 2. For example, a STAMP level 5 is adjacent to both a STAMP level 4 and a STAMP level 6. Therefore, if Rater 1 assigns a STAMP level 4 to a response and Rater 2 assigns a STAMP level 5 to that response, this would count towards this measure, since these two levels are adjacent to each other. Graham et al. (2012) suggest that when the rating scale has more than 5-7 rating levels, as is the case with the STAMP scale, exact + adjacent agreement should be close to 90%.

Quadratic weighted kappa (QWK)

Cohen's kappa, or κ, measures reliability between two raters by taking into account the possibility of agreement occurring by chance. For example, since the numerical STAMP scale in Writing and Speaking is a 9-point scale, going from STAMP level 0 to STAMP level 8, there is a 11.11% chance that any two raters would perfectly agree on a score simply by chance. At Avant, in addition to taking this chance agreement into account, we use quadratic weights when

calculating kappa, which means a higher penalty is assigned to scores that are farther away from each other. In other words, observing a difference between a STAMP level 3 and a STAMP level 7 between two ratings to the same response is more problematic than observing a difference between a STAMP level 3 and a STAMP level 4. Williamson et. al. (2012) recommend that QWK must be >= 0.70 and Fleiss (2003) notes that values above 0.75 show excellent agreement beyond chance for most purposes. A QWK value of 0 indicates agreement simply at the level of chance between two sets of ratings whereas a value of 1 indicates perfect agreement.

Standardized Mean Difference (SMD)

This measure shows the extent to which two raters may be using a rating scale in a similar way. It shows the difference of the mean of two sets of scores (*i.e*., Rater 1 vs. Rater 2) standardized by the pooled standard deviation of those two sets. Ideally, neither rater should prefer or avoid awarding levels at a certain point of a rating scale (for example, avoid giving either STAMP 0s or STAMP 8s). In other words, both raters should make equal use of the rating scale (STAMP 0 - STAMP 8) and the scores awarded should be dependent only on the level of proficiency shown in the response itself. It is recommended that the value for this measure should be <= 0.15 (Williamson et al., 2012), ensuring that the distribution of both sets of scores is acceptably similar.

Spearman's Rank-Order Correlation (ρ)

This measure indicates the strength of association between two variables, in this case the STAMP level assigned by Rater 1 and the STAMP level assigned by Rater 2. It is expected, if the team of raters are well trained and clearly understand the rating rubric, that whenever Rater 1 assigns a high proficiency level to a response, Rater 2 would also assign a high level. In other words, we expect the two sets of scores to move together (up or down) if the raters are indeed evaluating the same construct. We use Spearman's rank-order correlation coefficient instead of Pearson product-moment correlation since the former is preferred when the ratings are ordinal, as in the case of STAMP proficiency levels. A correlation coefficient of 0.80 or above is considered to be strong across various fields (Akoglu, 2018).

2 STAMP Levels Apart

This measure, expressed as a percentage, indicates the percentage of times in which two ratings to the same response have been observed to be 2 STAMP levels apart (for example, Rater 1 awards a STAMP level 4 to a response and Rater 2 awards a STAMP level 6).

Reliability and Accuracy of Scores by Avant Raters Across Various Languages

We now turn our attention to the quality of the ratings, in view of the statistics above, for the Writing and Speaking sections of STAMP 4S and STAMP WS across several representative languages. We provide below results based on two different sets of comparisons:

Rater 1 vs Rater 2

We compare the STAMP level awarded by Rater 1 to the STAMP level awarded by Rater 2 across a large number of responses in that language that were rated by at least two raters. This provides support for the *reliability* of the ratings provided by two randomly-assigned Avant raters. As previously mentioned, two raters could award the exact same STAMP level to an essay and both could still be incorrect in their rating, vis-a-vis what the actual rating should have been for that response. For that reason, we do not include exact agreement measures between Rater 1 and Rater 2. Instead, we focus on Exact + Adjacent Agreement and also report on accuracy measures between the score awarded by Rater 1 (who rates solo 80% of the time) and official scores (see below).

Rater 1 vs Official Score

In order to assess the *accuracy* of the levels assigned by Avant raters to responses, we look at a large number of instances in which a response was scored by two or more raters. We then compare the official score assigned to that response in the system (which is derived from the individual ratings for that response, as previously explained) to the score assigned by Rater 1 only. This provides us with an indication of how accurately a response is rated when only one Avant rater rates a response (which happens 80% of the time).

Tables 1 and 2 show the statistical measures for the Writing and Speaking sections of five representative STAMP 4S languages.

STAMP 4S Language Writing	Number of Responses in Dataset	Exact Agreement (Rater 1 vs. Official Score)	Exact + Adjacent Agreement Rater 1 vs. Rater 2 (Rater 1 vs. Official Score)	Quadratic Weight Kappa (QWK) Rater 1 vs. Rater 2 (Rater 1 vs. Official Score)	Standardized Mean Difference (SMD) Rater 1 vs. Rater 2 (Rater 1 vs. Official Score)	Spearman's Rank-Order Correlation (R) Rater 1 vs. Rater 2 (Rater 1 vs. Official Score)	2 STAMP Levels Apart Rater 1 vs. Rater 2 (Rater 1 vs. Official Score)
Arabic	*n* = 3,703	*(84.8%)*	96.78% *(98.62%)*	0.93 *(0.96)*	0.00 *(0.01)*	0.94 *(0.96)*	2.80% *(1.24%)*
Spanish	*n* = 4,758	*(84.15%)*	99.09% *(99.79%)*	0.91 *(0.95)*	0.00 *(0.00)*	0.90 *(0.95)*	0.90% *(0.20%)*
French	*n* = 4,785	*(83.66%)*	99.22% *(99.79%)*	0.91 *(0.95)*	0.00 *(0.00)*	0.91 *(0.95)*	0.77% *(0.20%)*
Chinese Simplified	*n* = 4,766	*(88.46%)*	99.79% *(99.91%)*	0.95 *(0.96)*	0.00 *(0.00)*	0.95 *(0.97)*	0.00% *(0.00%)*
Russian	*n* = 3,536	*(92.17%)*	99.71% *(99.88%)*	0.95 *(0.97)*	0.00 *(0.00)*	0.94 *(0.97)*	0.28% *(0.11%)*

Table 1. Rater Reliability and Accuracy Statistics for the Writing Section of Five Representative STAMP 4S Languages.

STAMP 4S Language Speaking	Number of Responses in Dataset	Exact Agreement (Rater 1 vs. Official Score)	Exact + Adjacent Agreement Rater 1 vs. Rater 2 (Rater 1 vs.	Quadratic Weight Kappa (QWK) Rater 1 vs. Rater 2 (Rater 1 vs.	Standardized Mean Difference (SMD) Rater 1 vs. Rater 2	Spearman's Rank-Order Correlation (R) Rater 1 vs. Rater 2 (Rater 1 vs. Official Score)	2 STAMP Levels Apart Rater 1 vs. Rater 2 (Rater 1 vs. Official Score)
Arabic	*n* = 3,363	*(84.96%)*	96.07% *(98.13%)*	0.92 *(0.95)*	-0.02 *(0.01)*	0.93 *(0.96)*	3.27% *(1.42%)*
Spanish	*n* = 4,078	*(80.37%)*	98.13% *(99.29%)*	0.92 *(0.96)*	0.00 *(0.00)*	0.91 *(0.95)*	1.74% *(0.00%)*
French	*n* = 4,530	*(80.19%)*	98.54% *(99.47%)*	0.91 *(0.95)*	-0.01 *(0.02)*	0.92 *(0.95)*	1.39% *(0.00%)*
Chinese Simplified	*n* = 4,651	*(82.24%)*	99.31% *(99.76%)*	0.94 *(0.95)*	0.00 *(0.00)*	0.94 *(0.96)*	0.00% *(0.00%)*
Russian	*n* = 3,392	*(88.30%)*	98.99% *(99.94%)*	0.92 *(0.96)*	-0.01 *(-0.01)*	0.91 *(0.95)*	1.01% *(0.00%)*

Table 2. Rater Reliability and Accuracy Statistics for the Speaking Section of Five Representative STAMP 4S Languages.

Tables 3 and 4 show the statistical measures for the Writing and Speaking sections of three representative STAMP WS languages.

STAMP WS Language Writing	**Number of Responses in Dataset**	**Exact Agreement** *(Rater 1 vs. Official Score)*	**Exact + Adjacent Agreement** *Rater 1 vs. Rater 2 (Rater 1 vs. Official Score)*	**Quadratic Weight Kappa (QWK)** *Rater 1 vs. Rater 2 (Rater 1 vs. Official Score)*	**Standardized Mean Difference (SMD)** *Rater 1 vs. Rater 2 (Rater 1 vs. Official Score)*	**Spearman's Rank-Order Correlation (R)** *Rater 1 vs. Rater 2 (Rater 1 vs. Official Score)*	**2 STAMP Levels Apart** *Rater 1 vs. Rater 2 (Rater 1 vs. Official Score)*
Amharic	*n = 209*	*(95.79%)*	99.52% *(100%)*	0.98 *(0.99)*	-0.01 *(0.00)*	0.98 *(0.99)*	0.00% *(0.00%)*
Haitian Creole	*n = 125*	*(94.69%)*	97.60% *(100%)*	0.97 *(0.99)*	0.02 *(-0.02)*	0.97 *(0.99)*	2.40% *(0.00%)*
Vietnamese	*n = 1,542*	*(94.38%)*	98.57% *(99.02%)*	0.96 *(0.97)*	-0.01 *(0.01)*	0.97 *(0.98)*	0.00% *(0.00%)*

Table 3. Rater Reliability and Accuracy Statistics for the Writing Section of Three Representative STAMP WS Languages.

STAMP WS Language Speaking	**Number of Responses in Dataset**	**Exact Agreement** *(Rater 1 vs. Official Score)*	**Exact + Adjacent Agreement** *Rater 1 vs. Rater 2 (Rater 1 vs. Official Score)*	**Quadratic Weight Kappa (QWK)** *Rater 1 vs. Rater 2 (Rater 1 vs. Official Score)*	**Standardized Mean Difference (SMD)** *Rater 1 vs. Rater 2 (Rater 1 vs. Official Score)*	**Spearman's Rank-Order Correlation (R)** *Rater 1 vs. Rater 2 (Rater 1 vs. Official Score)*	**2 STAMP Levels Apart** *Rater 1 vs. Rater 2 (Rater 1 vs. Official Score)*
Amharic	*n = 225*	*(96.21%)*	100% *(100%)*	0.99 *(0.99)*	0.00 *(0.00)*	0.99 *(0.99)*	0.00% *(0.00%)*
Haitian Creole	*n = 132*	*(97.91%)*	100% *(100%)*	0.99 *(0.99)*	0.00 *(0.00)*	0.99 *(0.99)*	0.00% *(0.00%)*
Vietnamese	*n = 1,180*	*(97.01%)*	99.83% *(99.83%)*	0.99 *(0.98)*	0.00 *(0.01)*	0.98 *(0.99)*	0.00% *(0.00%)*

Table 4. Rater Reliability and Accuracy Statistics for the Speaking Section of Three Representative STAMP WS Languages.

Discussion

A high level of reliability and accuracy is fundamental to the validity of test scores and their intended uses. What is deemed minimally acceptable in terms of reliability and accuracy will however depend on the specific field (medicine, law, sports, forensics, language testing, etc), as well as on the consequences of awarding an inaccurate level to a specific examinee's set of responses, and on the rating scale itself. For example, agreement will tend to be lower the higher the number of categories available in a rating scale. In other words, more disagreement between any two raters can be expected if they must assign one of ten possible levels to a response than if they must assign one of only four possible levels.

The statistics seen above for the Writing and Speaking sections of both STAMP 4S and STAMP WS show a high level of both reliability (Rater 1 vs. Rater 2 scores) and accuracy (Rater 1 vs. Official Scores). Of the eight languages evaluated, the reliability seen by Exact + Adjacent Agreement between Rater 1 and Rater 2 is always at a minimum (and often considerably higher) of 96.78% for Writing and 96.07% for Speaking. Additionally, cases in which the ratings by two raters were more than two STAMP levels apart were very seldom observed. The level of accuracy for all eight languages, seen by the Exact Agreement statistics between Rater 1's score and the Official score for each response is always at a minimum of 83.66% (but often considerably higher) for Writing and 80.19% for Speaking, with Exact + Adjacent Agreement always at a minimum of 98.62% for Writing and 98.13% for Speaking. The values for Quadratic Weighted Kappa (QWK) show a very high level of agreement between both Rater 1 vs. Rater 2

and between Rater 1 vs. Official Scores, while the correlation between Rater 1 and Rater 2 scores, as well as between Rater 1 and Official Scores,have been shown to be very high. Finally, the SMD (Standardized Mean Differences) coefficients show that the STAMP scale is being used in a very similar fashion by Avant raters.

The statistics above provide evidence of the high quality of the rater selection and training program at Avant Assessment and of our methodology in identifying operational raters who may need to be temporarily removed from the pool of raters and given targeted training. It shows that when any two raters may differ in the STAMP level assigned to a response, the difference will rarely be of more than 1 STAMP level, with both raters assigning the exact same level in the great majority of cases. Coupled with the fact that an examinee's final, official score in either the Writing or Speaking section of STAMP is based on their individual STAMP scores across three independent prompts, the results herein provide strong evidence that an examinee's final score for the Writing and Speaking sections of STAMP can be trusted to be a reliable and accurate representation of their level of language proficiency in these two domains.

REFERENCES

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish journal of emergency medicine*, *18*(3), 91-93.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.

Feldt, L. S., & Brennan, R. (1989). Reliability. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 105–146). New York: Macmillan.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). Statistical methods for rates and proportions. 3rd ed. Wiley.

Graham, M., Milanowski, A., & Miller, J. (2012). Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings.

Matrix Education (2022). Physics Practical Skills Part 2: Validity, Reliability and Accuracy of Experiments. Retrieved on August 11, 2022 (click here to go to source).

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational measurement: issues and practice*, *31*(1), 2-13.