Avant STAMP Test Maintenance

Victor D. O. Santos, Ph.D.

Director of Assessment and Research

Avant Assessment, LLC

2/12/2024

1. Test Overview

The STAMP 4S test is a computer-adaptive, multistage assessment of language proficiency, currently available in over 47 languages. STAMP 4S assesses proficiency across the four language domains of Reading, Writing, Listening, and Speaking. The Reading and Listening sections are multiple-choice and are automatically scored, whereas the Writing and Speaking sections, comprised of constructed responses, are scored by Avant raters trained on the STAMP proficiency scale, which is based on the ACTFL Proficiency scale.

Instead of employing a design in which all test-takers see the exact same items during test administration (thus reducing the reliability of the scores), the Reading and Listening sections of STAMP 4S employ a multistage adaptive design in which the difficulty of the items adapts to the level of proficiency a test-taker shows throughout the test. Not only does this make for a better experience for test-takers but it also helps increase the accuracy of measurement when compared to a linear, non-adaptive test format. Test security is also increased through an adaptive format since different test-takers will encounter different items in the test, thus reducing item exposure. The design of the Reading and Listening sections of STAMP 4S is shown in Figure 1.

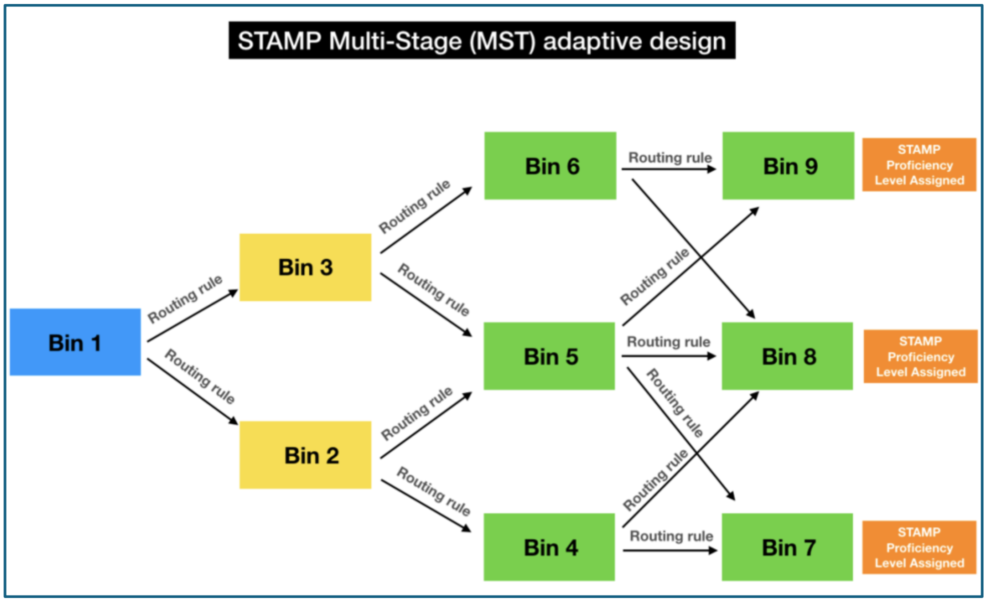
STAMP Multi-Stage (MST) adaptive design
Bin 1
Bin 2
Bin 3
Bin 4
Bin 5
Bin 6
Bin 7
Bin 8
Bin 9
Routing rule
STAMP Proficiency Level Assigned
STAMP Proficiency Level Assigned
STAMP Proficiency Level Assigned

Figure 1. *The multistage adaptive design employed in STAMP 4S.*

In the Writing and Speaking sections, test-takers must type (in the case of Writing) or record (in the case of Speaking) their responses to three level-appropriate prompts/scenarios. The specific scenarios/prompts a test-taker receives is determined by their STAMP score in the Reading and Listening sections and on a randomization algorithm built into the STAMP system. The pool of prompts is larger than the number of prompts that must be delivered during the test so as to ensure that test-takers do not all receive the same prompts, once again reducing item exposure, which could negatively affect test scores.

The development of any STAMP test follows the same ten-step process, outlined in the white paper *The Development of a STAMP Test: Support for Test Validity*. Once a STAMP test has been developed, ongoing test maintenance is performed to continue to ensure the accuracy, reliability, and validity of test scores. In the following sections we outline the maintenance steps the STAMP test adheres to.

2. Overview of Maintenance Plan

2.1. Writing and Speaking Sections

The Writing and Speaking sections of a STAMP 4S test are completely refreshed on a yearly basis. In December each year, all prompts in both these productive sections are retired and replaced by a new set of prompts, which will be active for the subsequent twelve months, until they are in turn replaced by a new set of prompts. The need to refresh all prompts in the Writing and Speaking sections yearly stems from the fact that the pool of available prompts is smaller than the pool of available items in the Reading and Listening sections, which means that any one prompt will tend to have more exposure within a year than most items in the receptive Reading and Listening sections.

In addition to replacing all prompts yearly, the Rating and Research teams at Avant also run various statistical analyses throughout the year to ensure that Avant raters are grading responses to the expected standard. If a specific rater is not grading to expectations, they are immediately removed from the pool of raters, given additional training, and brought back to rate again only once they have shown they have met the accuracy expectations. This ensures the continuation of the high level of accuracy and reliability observed in the rating of responses in the Writing and Speaking sections, described in the white paper *Reliability and Accuracy of Ratings in the Writing and Speaking Sections of STAMP Tests.*

2.2. Reading and Listening Sections

The average item in the Reading and Listening sections tends to have less overall exposure than prompts in the productive sections due to the multistage adaptive design employed in the Reading and Listening sections of STAMP (Figure 1 above). Therefore, it is not necessary to replace all items in the pool every year, as is done with prompts in the Writing and Speaking sections. Overall, approximately 20-25% of the items are replaced yearly in these sections.

Items in the Reading and Listening sections are replaced according to the principles below: **Item Exposure**: The higher the number of people who have taken a specific item, the more urgent the need to replace that item with a brand new one. This is done to ensure item exposure (which could also lead to item leaking) does not impact test scores. Even with strict proctoring policies in place for administration of STAMP, the chance of an item leaking can never be said to be zero, for any test. Therefore, items with higher exposure rates have more priority for replacement than items with lower exposure rates, all things equal. Even if an item has not been leaked, we would not want a test-taker to take an item many times, across different test administrations.

Item Review: The content of each item is reviewed yearly to ensure that it stays relevant and fair to all groups of test-takers. If an item is shown to deviate from this, the item also becomes higher priority for revision or replacement.

Item Statistics: As part of Avant's yearly maintenance plan, items in the Reading and Listening sections of a STAMP test are scrutinized not only in terms of content but also in terms of their statistical properties. STAMP items are analyzed both within a Classical Test Theory (CTT) approach and an item-response theory (IRT) approach. It is unusual that STAMP items show undesirable statistics given the thorough test development process employed for STAMP tests (described in the white paper The Development of a STAMP Test: Support for Test Validity), but in the rare event when this does happen – which can be due to several reasons – items with less-than-ideal statistical qualities are also marked as high priority for replacement.

It must be noted that item replacement is one of the test maintenance strategies employed at Avant. Another strategy is what Avant refers to as a *reshuffle*, in which the position of an item in the test, within the multistage adaptive design, changes. For example, if an item that has been assigned to Bin 1 in the test (see Figure 1 above) has already been taken by many people, the item can then be assigned to a different bin (for example, bin 8) to reduce item exposure.

Another advantage of this reshuffle strategy is that it also allows items that have been taken by a smaller number of test-takers (for example, items in bin 7) to be moved to earlier bins to increase the number of people who have taken the item. Although having an item taken by too many test-takers is problematic due to high exposure, it is also important to ensure that enough test-takers encounter an item, otherwise it is not possible to run reliable statistics on that item. The reshuffle strategy also increases the chance that any test-taker will encounter different items on the test based on a pre- and post-reshuffle, even if their proficiency has not changed. Given the fact that Avant sets its cut-scores on Rasch values and not on number of correct answers, the scoring algorithm automatically takes into account any eventual change in the difficulty of the items along a specific adaptive path due to item replacement or reshuffle.

2.3. Live Neutral Test Items

The Test Development team at Avant is always developing new items for our STAMP tests to ensure the feasibility of the maintenance strategy described above. Until recently, Avant field tested newly developed STAMP items in a non-live test and with a representative population of test-takers. Although that is a commonly used approach, we have more recently started collecting information on new items by inserting them into a live STAMP test but not awarding any points to them.

The advantage of adding newly developed items as neutral items in a live test is that test-takers do not know which items are scored and which are not, meaning they will give their best on all items. Additionally, this neutral-item approach allows the Research and the Content Development teams at Avant to collect statistics on new items much more quickly and easily than would be the case if a separate field test were needed. Once the statistics (including difficulty) of a newly developed item have been collected and verified to be desirable, the item can then be added to a live STAMP test and have points assigned to it (*i.e.*, become a scored item).

3. Conclusion

A test refresh plan is a crucial component of a healthy and high-quality standardized language proficiency test. A refresh plan helps ensure that the items on the test are behaving as expected and that factors unrelated to language proficiency are not significantly affecting test scores. Without such a plan, test scores run the risk of no longer having the same meaning and interpretation that they originally did.

By yearly refreshing all the prompts in the Writing and Speaking sections of STAMP and by considering test-taker numbers, potential item exposure, and item statistics in its decisions regarding which items to prioritize for refreshing in the Reading and Listening sections of STAMP, Avant Assessment continues to ensure the high-quality of STAMP and the meaning and interpretations of STAMP scores.