The Development of a STAMP Test: Support for Test Validity

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.

The Development of a STAMP Test:
Support for Test Validity

Victor D.O. Santos, PhD
Director of Assessment and Research
Avant Assessment
November 1, 2019

The development of any high-quality test is a significant endeavor. This is especially the case with a computerized-adaptive test such as the STAMP (**STA**ndards-Based **M**easurement of **P**roficiency), dealing with the construct of language proficiency. The STAMP test assesses test takers' proficiency across all four language domains: Reading, Writing, Listening, and Speaking. The proficiency scores on the STAMP test are on a scale of 1 (Novice-Low) to 9 (Advanced-High), with the STAMP scale being aligned with the ACTFL proficiency scale, as seen in Figure 1:

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

Figure 1*. Alignment of the STAMP scale with the ACTFL scale.*

The development of a STAMP test is a multi-step process that takes, on average, six months to complete, from the development of the technical requirements and specifications for the test to the official release of the test to the public. Partial support for the validity of an assessment, given its intended uses, comes precisely from the processes employed in test development. The attention given to these processes in the development of STAMP is reflected, for example, in the Avant STAMP Annual Averages, in which we see the expected increase in language proficiency as students/test takers move into higher grades and as they have been studying the language for a longer time.

The ten general steps involved in the development of the Reading and Listening sections of a STAMP test can be seen in Figure 2:

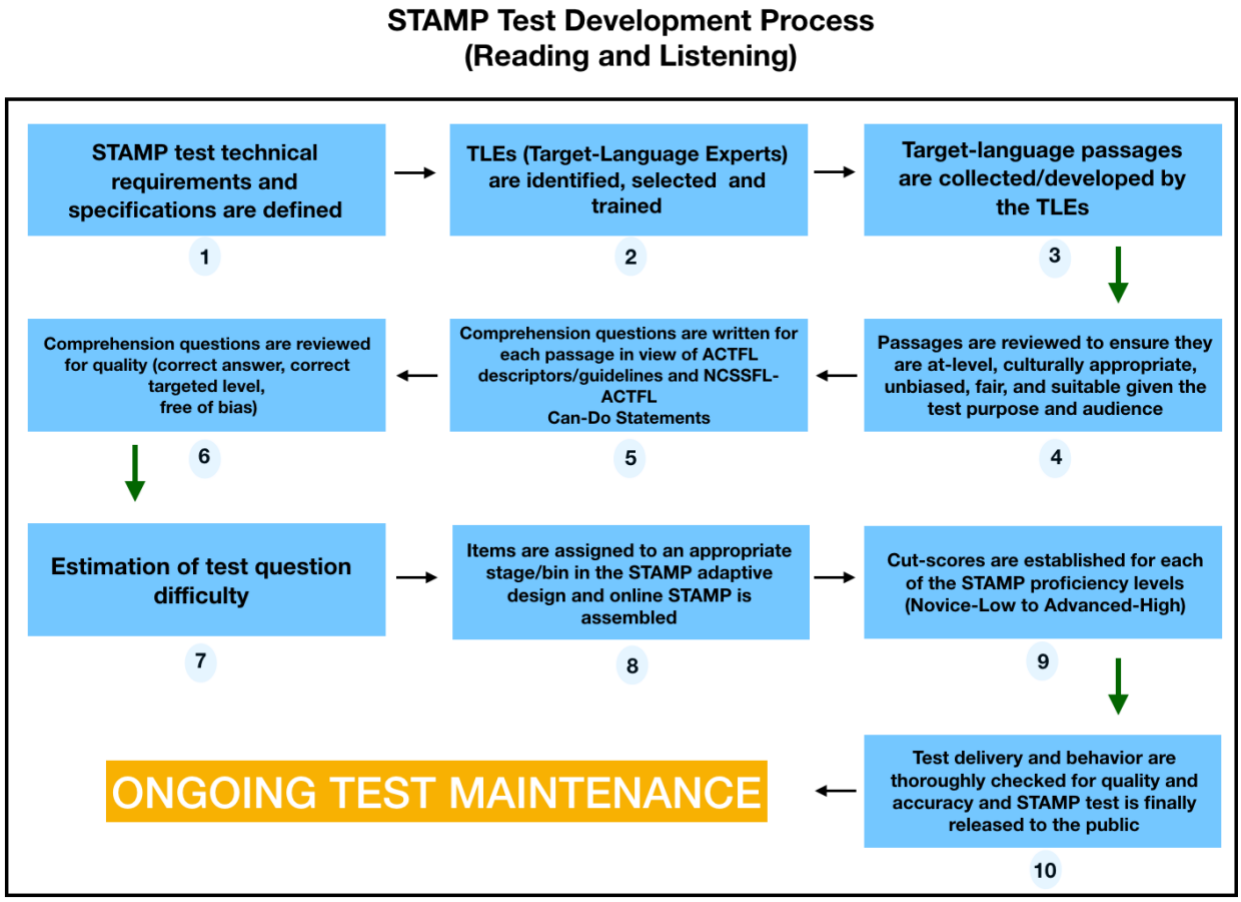


Figure 2. *The ten general steps involved in the development of a STAMP test.*

In what follows, a more detailed, non-technical discussion of what each of the ten steps involves is provided.

1. STAMP Test Technical Requirements and Specifications are Defined

In order for a defensible and valid assessment such as the STAMP to be developed, it is important that test specifications be defined at the very beginning of test development, prior to any passages being written/collected and prior to any items being written.

As Davidson (2012) notes, "a test specification ("spec") is a generative blueprint for test items or tasks from which many equivalent test items or tasks can be produced" (p.4). Basically, a test specification is to a test what a recipe is to a cake; it allows for reproducible assessments (which includes content and questions) and contributes to the reliability and validity of the assessment.

During the definition of the test specifications, Avant test developers establish, among other things, the following guidelines or specifications:

(a) the general description of the test (including format and purpose);
(b) the attributes/characteristics of the items (passages + questions) in the test;
(c) the response attributes[1] that are required for item writing and for test takers to obtain various scores on an item;
(d) the distribution of the various possible item types[2] in the STAMP test (*e.g.*, zone click, picture select, multiple choice, multi-select)
(e) the scope of topics for the various passages at the different major levels of proficiency (Novice, Intermediate, and Advanced)
(f) the number of items at each of the major proficiency levels that will be presented at each stage of the (adaptive) test.

In the Appendix, some of the item types that may be present on a STAMP test can be seen. For the actual item types that will be present in a given STAMP test, readers can check the Avant website for sample STAMP tests in all currently available languages.

2. TLEs (Target-Language Experts) are Identified, Selected, and Trained

With over a dozen languages in which the STAMP 4S/4Se test is available, it is crucial that Avant have at its disposal a highly qualified team of TLEs (target-language experts). Avant TLEs are well-educated native speakers of one of the languages we offer the STAMP test in. Once TLEs are identified, their credentials and suitability for the job are checked. If selected to join the team working on the development of a given STAMP test, the TLEs then receive training by Avant's testing experts on the ACTFL proficiency scale (many of our TLEs are already familiar with it), on the construct of the STAMP test(s), and on the requirements and specifications for the test.

3. Target-Language Passages are Collected/Developed by the TLEs

Once they have been fully trained and are deemed ready to start work on a STAMP development project, TLEs start collecting and, on occasion, developing[3] reading and listening passages in the target language that are appropriate for a certain proficiency level in terms of length, topic, text type/genre, linguistic complexity, and opportunity for assessment points (i.e., for good questions to be written on). Each passage must closely follow the requirements in the test specifications.

1 An example of a response attribute (RA) would be "*test takers must select the best answer among four possible options".*

2 Not all item types may appear on a given STAMP test.

3 Especially at the lower levels of proficiency, it may not always be possible to employ authentic passages (collected in the real world). Therefore, on occasion, TLEs may be asked to develop a passage that fits the requirements, while still being natural and representative of the level, culture, and language at hand.

4. Passages are Reviewed to Ensure they are At-Level, Culturally Appropriate, Unbiased, Fair, and Suitable Given the Test Purpose and Audience

Every passage that finds its way into a STAMP test must be at the appropriate proficiency level (Novice, Intermediate, or Advanced), culturally appropriate, free of bias, fair to all test takers, and suitable given the purpose and audience of the STAMP test in question.

It is important that the proficiency level of the passages be clearly and accurately identified, especially given the adaptive nature of the STAMP test, in which the difficulty of the items adapts to the proficiency level of the test taker so as to increase accuracy of measurement and improve the test-taking experience. This will later on make it easier for comprehension questions to be written at the right level for those passages and to assign the passage + question pair to the appropriate stage of the STAMP adaptive algorithm.

It is also important that any passage selected for inclusion in the Reading or Listening section of a STAMP test be culturally appropriate and relevant. Additionally, passages must be both fair and free of bias. Fairness concerns the fact that the topic and content of a passage must *not* be upsetting or offensive to any groups of test takers, and to the fact that all possible groups of people (whether based on gender, country of origin, race, religion, or other factors) must be treated and portrayed with respect. Being free of bias, on the other hand, "refers to construct-irrelevant [*i.e.*, invalid] components that result in systematically lower or higher scores for identifiable groups of examinees (AERA, APA, & NCME, 2014). For example, if it is deemed that test takers with prior knowledge of the topic of a passage have a higher chance of getting a question associated with that passage correct, then that passage would not be suitable for inclusion in the STAMP test.
Finally, the passage must be suitable given the test purpose and audience of the STAMP test. The purpose of a STAMP test is to accurately measure one's language *proficiency* across all four language domains, and scores on our test should only depend on a test taker's language proficiency and nothing else. For example, topics in the STAMP 4Se must be more familiar, day-to-day topics appropriate for test takers of elementary school age, whereas topics for our STAMP 4S test are broader, given the audience of the latter (middle-school to adult test takers).

5. Comprehension Questions are Written for Each Passage in View of ACTFL Descriptors/Guidelines and NCSSFL-ACTFL Can-Do Statements

Once passages have been collected and successfully passed review, experienced, professional items writers at Avant begin writing comprehension questions for each of the passages in the test, in view of the targeted level of the passage, the ACTFL descriptors for language learners (ACTFL, 2012), the NCSSFL-ACTFL Can-Do statements (NCSSFL-ACTFL, 2017) and the ACTFL proficiency guidelines (ACTFL, 2012).

For the majority of Novice level passages, a single comprehension question is usually written. For Intermediate level passages, one or two questions may be written, and for Advanced passages, usually

at least two questions are written. Advanced passages are considerably longer and writing two or more questions for each Advanced passage allows for testing of more assessment points and for KSAs (Knowledge, Skills, and Abilities) to be assessed more efficiently, without requiring that test takers read a different passage each time they must answer a comprehension question.

6. Comprehension Questions are Reviewed for Quality (Correct Answer, Correct Targeted Level, Free of Bias)

One of the premises of test validity is that the test is scored correctly. What this means in the case of the STAMP test is that there should be only one correct answer that leads to a full score on a given question. For a multiple-choice question, for example, having the system accept an incorrect answer as correct due to a programming error would jeopardize the validity of the test. For that reason, every single comprehension question developed for a STAMP test goes through several levels of internal review to ensure that only the intended correct answer is accepted by the system and positively contributes to a test taker's score.

Prior to assembling a STAMP test, Avant test developers also ensure that each item is written to the correct intended level. In other words, if the item is intended to target the Intermediate proficiency level and to provide evidence of Intermediate-level proficiency, it is vital that there is sufficient internal agreement that this is the case. A comprehension question targeting the Intermediate level, for instance, is written to elicit evidence of Intermediate-level processing of the passage (whether written or spoken) and requires the application of grammatical, lexical and text-organizational knowledge, in addition to other types of knowledge, typical of the Intermediate level of proficiency.

Lastly, just as with passages, items should be free of bias. Answering a question correctly should depend solely on comprehension of the passage, and not on any previous knowledge a test taker may have that could give him or her an advantage over other candidates at the same level of language proficiency.

7. Estimation of Test Question Difficulty

In order to establish the difficulty of items that have been written for a given STAMP test, there are two methods commonly employed at Avant: *(a)* TLEs rate the difficulty of each item in a session via a modified Angoff panel (Cizek & Bunch, 2007) or *(b)* the entire pool of items written for the upcoming STAMP assessment is field tested on (administered to) hundreds of actual test takers representative of the test-taking population of interest in terms of age and language proficiency. The preference at Avant is to employ method (b), since it allows items to be calibrated through IRT[4] (Item-Response Theory) right after the field test and added to the item pool with its actual, observed item difficulty. In cases when this is logistically not viable, as is sometimes the case with LCTLs (less commonly taught/tested languages), method A above must be used instead as a temporary proxy for

[4] IRT is a modern measurement approach, with certain advantages over the Classical Test Theory (CTT) approach. At Avant, both are used in a complementary fashion.

item difficulty for test assembly purposes. If method A ends up being employed prior to the release of the STAMP test, actual/observed item difficulties will be calculated once the test has been in operation long enough for hundreds of test takers to have taken each item. In the end, the result is the same, with a given proficiency level (*e.g.*, Intermediate) and sub-levels (*e.g.*, Intermediate-Mid) being anchored to a particular IRT value for subsequent test maintenance.

8. Items are Assigned to an Appropriate Stage/Bin in the STAMP Adaptive Design and Online STAMP is Assembled

Once the difficulty of the items on a STAMP test is estimated through either method A or method B described in #7 above, the items can be assembled into the online STAMP platform for subsequent quality check and delivery to test takers.

At Avant, our STAMP assessments are computer-delivered, employing an adaptive design that is more modern and psychometrically robust than linear tests. In linear tests, a fixed (*a.k.a*. linear) form is developed and usually taken by all test takers, regardless of each test taker's individual language proficiency. The design employed for the Avant STAMP test is called *multistage adaptive test design* (MST). MST is a balanced compromise between linear test forms and item-level computer adaptive testing (Zheng, Nozawa, Gao, & Chang, 2012). Another advantage of multistage adaptive testing (MST) is that it allows for "more efficient and precise measurement across the proficiency scale" (Hendrickson, 2007).

In an MST design test, items are grouped into bins, with each bin containing items that target one or more levels of proficiency. As individual test takers progress through the STAMP test, they are routed (depending on the level of proficiency demonstrated so far on the test) into a bin that contains items that are more appropriate for their demonstrated level of proficiency. This allows for both better precision of measurement and for a better test-taking experience, since test takers, contrary to what happens with a linear test, will not come across many items that are either too easy or too difficult for them. Figure 3 shows the MST design employed in the STAMP assessments:

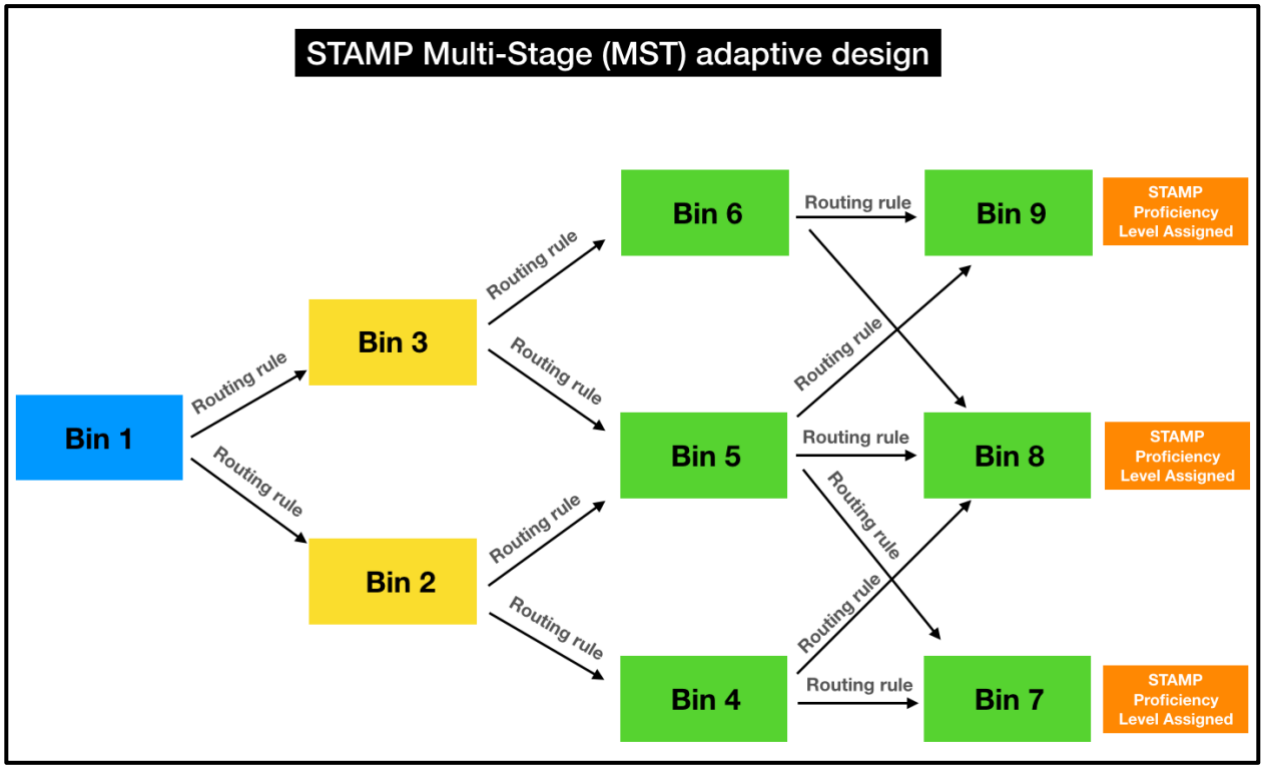


Figure 3. *STAMP multistage testing (MST) design.*

Once the Avant test development team has finished assigning each of the STAMP items to a certain bin in view of each item's difficulty and content properties (topics, assessment focus, etc), routing rules can be established for routing test takers into the various bins/stages in the test.

9. Cut-Scores are Established for each of the STAMP proficiency levels (Novice-Low to Advanced-High)

Once the assignment of items to bins and the routing rules have been established, proficiency levels can finally be reported for any test taker, based on the specific items encountered during the test, the specific path followed by each test-taker in the MST design, and the level of language proficiency demonstrated in their answers to the questions. There are several possible ways to define what evidence of proficiency at each of the proficiency levels looks like on the STAMP test, one of which is by assigning an IRT-based cut-score to each of the possible STAMP levels 1-9. This approach has the advantage of allowing test items to be replaced in our regular STAMP refresh process without affecting current and future determinations of proficiency levels.

10. Test Delivery and Behavior are Thoroughly Checked for Quality and Accuracy, and STAMP Test is Finally Released to the Public

Now that the entire online STAMP test has been assembled, with routing rules and cut-scores established, the Avant team starts its quality control (QC) procedures with regard to test delivery. We scrutinize every single detail of each item and take the test over and over again to ensure that it is routing test takers through the correct test paths and assigning an accurate final score/proficiency level. Once it is determined that the test is working optimally and that the routing, scores, and proficiency classifications are correct, the test is finally made available to the public.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.) (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. SAGE Publications Ltd.

Davidson, F. (2012). Test specifications. In C.A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics.* Blackwell Publishing.

Hendrickson, A. (2007). An NCME instructional module on multistage testing. Educational Measurement: Issues and Practice, *26*(2), 44-52.

Zheng, Y., Nozawa, Y., Gao, X., & Chang, H. H. (2012). Multistage Adaptive Testing for a Large-Scale Classification Test: Design, Heuristic Assembly, and Comparison with Other Testing Modes. ACT Research Report Series, 2012 (6). *ACT, Inc.*

Glossary

Angoff panel

An Angoff panel consists of a group of experts in the construct of the test (in the case of the STAMP test, proficiency in a specific target language). These would normally include individuals who have experience teaching that language to students at various proficiency levels as well as native speakers of the language familiar with the ACTFL proficiency scale. The group of experts is convened into a panel whose main task is to estimate how test takers at various levels of proficiency would fare on the STAMP test, given the items included on the test.

Bin

A bin in the STAMP adaptive design refers to a set of items that are grouped together according to their level of difficulty. As test-takers progress through the STAMP test, they are routed to different bins (see Figure 3), depending on how well they have done on the test up to that point in time. The main purpose of the bins is to increase measurement accuracy, given that test-takers will encounter fewer and fewer items that are either too hard or too easy for them as they move through the STAMP adaptive algorithm.

Cut-scores

Cut-scores are a mathematical operationalization of proficiency levels. In order to assign a proficiency level to a test-taker at the end of their STAMP administration, the STAMP system needs to know which items the test-taker encountered during the test (in other words, which *bins* they "visited") as well as how many of those items they answered correctly or incorrectly. Once that information is known, we can compare it against a table of cut-scores, which defines the number of correct items needed for test-takers to be classified as Novice-Low, Novice-Mid, Novice-High, Intermediate-Low, and so on, given the bins they were routed to during the test.

Field testing

Field testing refers to a practice that involves administering test items (or an entire test) to a large number of test-takers prior to releasing those items or entire test to the public in operational form. The purpose of the field test is to ensure that the items (and/or test) are working as expected.

Stage

In the STAMP adaptive design, a stage refers to a routing point in the STAMP system. Figure 3, for example, shows four stages:

Stage 1 (bin 1): all test-takers are routed to bin 1 when they start the STAMP test.

Stage 2 (bin 2 or 3): once test-takers answer the questions in bin 1, they will be routed to either bin 2 OR bin 3, depending on their score in bin 1.

Stage 3 (bin 4, 5, or 6): once test-takers answer the questions in bin 2 OR bin 3, they will be routed to either bin 4, bin 5, OR bin 6, depending on their score in the previous bins they encountered.

Stage 4 (bin 7, 8, or 9): once test-takers answer the questions they encountered in stages 1,2, and 3, they will be routed to either bin 7, bin 8, OR bin 9, depending on their score in the previous bins.

APPENDIX

Sample Avant STAMP item types
Zone click & Picture select

STAMP Portuguese 4S
Novice level (item type: zone click)

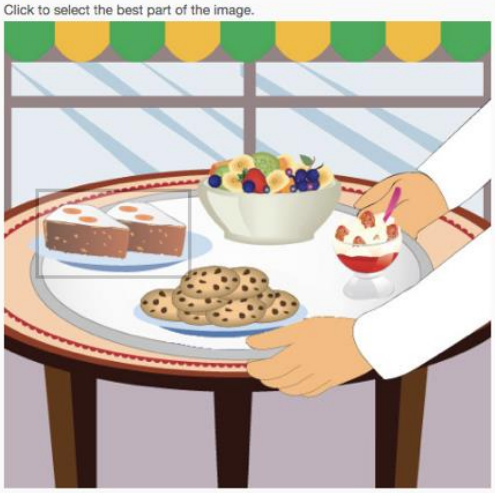
Sample Listening Item – Novice Level

Situation

You are at a local restaurant and overhear the following conversation.

What is the man ordering?

Click to select the best part of the image.



STAMP Hebrew 4S
Novice level (item type: picture select)

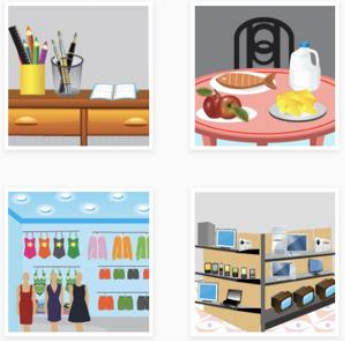
Sample Reading Item – Novice Level

Situation

You receive this flyer in the mail.

What things are included in the sale?

Choose the best image.



Sample Avant STAMP item types
Multiple choice & Multi-Select

STAMP German 4S
Intermediate level (item type: multiple choice)

Sample Reading Item – Intermediate Level

Situation

You see this ad in a local newspaper.

Lieferdienst Riegel - einfach lecker!

Gehen Sie doch mal bei sich zuhause essen! Wir liefern kostenlos und schnell: warme Speisen, frische Salate und wunderbare Nachspeisen. Deutsche und internationale Spezialitäten.

Montag ist Kuchentag! Alle Kuchen- und Tortenstücke 2,50 € ab 18 Uhr. Täglich von 12 Uhr bis 22 Uhr und Freitags und Samstags bis 24 Uhr. Rufen Sie uns an unter **0761-2252860** oder erreichen Sie uns im Internet unter **www.lieferdienst-riegel.de.**

What does this business offer?

Choose the best answer.

- meals delivered to your home
- international specialty groceries
- cooking classes for seniors
- custom-decorated birthday cakes

STAMP Portuguese 4S
Advanced level (item type: multi-select)

Sample Reading Item – Advanced Level

Situation

You are reading a story that a classmate wrote for a class assignment.

Há várias semanas, Isabela vinha aguardando ansiosamente pelas próximas férias de acampamento com seus pais. Ela já havia acampado diversas vezes em Leme, no interior de São Paulo, onde há um lindo lago de águas claras, árvores exuberantes, trilhas e uma lojinha de esquina onde sempre compravam suprimentos e guloseimas.

Finalmente, a véspera das férias! Naquela manhã, seu pai iria buscar o trailer que sempre alugava. No momento em que Isabela entrou na cozinha para tomar seu café da manhã, a menina notou os rostos tristes de seus pais. "O que foi?", perguntou Isabela. "Vocês não estão empolgados que vamos sair de férias?"

"Isabela", respondeu sua mãe. "O patrão do seu pai acabou de ligar dizendo que um dos engenheiros mecânicos da empresa pediu demissão e que seu pai vai ter que cobri-lo". Infelizmente nós vamos ter que adiar nossa viagem por duas semanas.

Isabela ficou com muita raiva. Após dar uma meia-volta, correu em direção ao seu quarto e bateu a porta com força. De dentro do quarto, Isabela ouviu seu pai ligando o carro para ir para o trabalho. A menina se trancou no quarto o dia inteiro.

À noite, ao chegar do trabalho, o pai de Isabela se aproximou do quarto da filha. "Eu sei que você está bastante decepcionada, minha filha", lhe disse seu pai. "Mas olha só o que eu trouxe para você: uma barraca só sua! Você já pode começar a treinar como montar a barraca no quintal e daqui a duas semanas pode levá-la para o acampamento e dormir nela. E mais uma coisa: a sua prima predileta também virá com a gente desta vez. O que você acha?". Isabela se encheu de alegria e deu um abraço bem forte em seu pai. Ela mal podia esperar a hora de saírem de férias.

What helped Isabel get over her disappointment?

Choose the two best answers.

- Her father bought her a new tent.
- She invited a friend to come for a visit.
- Her parents took her on a hike.
- She went on a camping trip by herself.
- Her cousin agreed to go camping.

Avant

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

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