INSTITUTE FOR TEST RESEARCH
AND TEST DEVELOPMENT

RESEARCH PAPERS IN ASSESSMENT

Erwin Tschirner, Olaf Bärenfänger (eds.)

Erwin Tschirner

Examining the Validity and Reliability of the ITT Vocabulary Size Tests

Volume 3

Bibliographische Informationen der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet über http://dnb.ddb.de abrufbar.

Die "Research Papers in Assessment" sind eine Reihe des Instituts für Testforschung und Testentwicklung e. V. (ITT), in der Forschungsergebnisse, Tagungsbeiträge und wichtige Einzeldarstellungen veröffentlicht werden.

Institut für Testforschung und Testentwicklung e.V. Leipzig
c/o Herder-Institut
Universität Leipzig
Beethovenstraße 15
04107 Leipzig
www.itt-leipzig.de

Herausgeber:
Erwin Tschirner, Universität Leipzig
Olaf Bärenfänger, Universität Leipzig

Format und Layout:
Nadja Nitsche

URN des Bandes: http://nbn-resolving.de/urn:nbn:de:bsz:15-qucosa2-760957
URN der Reihe: http://nbn-resolving.de/urn:nbn:de:bsz:15-qucosa-188813

ISSN 2366-6870

Vorwort der Herausgeber

Das aussagekräftige Messen und Bewerten von Fremdsprachenkenntnissen gewinnt zunehmend an Bedeutung. Sowohl in den Bereichen Beruf und Bildung, aber auch im Privaten und nicht zuletzt im Zuge erheblicher Zu- und Abwanderungsbewegungen weltweit spielt das Beherrschen, Fördern und Evaluieren von Sprachen eine maßgebliche Rolle. Die Reihe *Research Papers in Assessment*, herausgegeben vom Vorstand des Instituts für Testforschung und Testentwicklung e. V., präsentiert aktuelle Studien zur validen und reliablen Messung von Sprachkenntnissen, zu High- und Low-Stakes-Tests, zu Testkonzepten für Unterricht und Lehrmaterialien, zu diagnostischen Testverfahren und damit verbundener individueller Sprachförderung, Sprachbedarfsanalysen und allen damit verbundenen Themen. Die Reihe erscheint als Online-Publikation, um aktuelle Forschungsergebnisse möglichst rasch interessierten WissenschaftlerInnen, Lehrkräften und mit dem Testen von Fremdsprachenkenntnissen betrauten Institutionen zugänglich zu machen und diese in die Testpraxis umsetzen zu können.

Die Herausgeber

Erwin Tschirner

Olaf Bärenfänger

Examining the Validity and Reliability of the ITT Vocabulary Size Tests

Erwin Tschirner

Contents

List of Tables

Introduction

The Institute for Test Research and Test Development (ITT) has provided complimentary Vocabulary Size Tests (VST) in 15 languages to language learners and their teachers, measuring their own or their learners' receptive and productive vocabulary sizes (Institute for Test Research and Test Development, n. d.). Receptive vocabulary size, in particular, correlates highly with reading and listening proficiency levels and provides a measure of overall language proficiency (Laufer & Nation, 2012; Milton, 2009; Qian & Lin, 2019; Schmitt, 2008; Stæhr, 2008). A study done by Hacking, Rubio, & Tschirner (2019), e. g., documented very high correlations between reading proficiency and receptive vocabulary size (Chinese: $r = 0.84$; Russian: $r = 0.87$; Spanish: $r = 0.88$), while Stæhr (2008) found high correlations between a receptive vocabulary size test and English reading ($r = 0.83$), listening ($r = 0.69$), and writing proficiency ($r = 0.73$). A vocabulary size of 2,000 words is generally associated with a reading proficiency of A2, while vocabulary sizes of 3,000 and 5,000 words are associated with B1 and C1, respectively (Milton, 2010; Huhta et al., 2011; Tschirner, 2019).

The VST is modeled after the English Vocabulary Levels Test pioneered by Paul Nation (Nation, 1990). The VST measures how many of the most frequent 5,000 words of a language are known. It uses the word frequency lists of the Routledge Frequency Dictionaries (Routledge, n. d.), which contain the

most frequent 5,000 words of a language. The VST consists of five bands: the most frequent 1,000; 1,001 to 2,000; 2,001 to 3,000; 3,001 to 4,000; and 4,001 to 5,000 words. There are two kinds of tests: a receptive VST, which measures the size of someone's sight vocabulary, i. e., the words that are understood when reading, and a productive one, which measures the size of someone's productive vocabulary, i. e., the words one is able to say or write to express a particular meaning.

The receptive vocabulary size test consists of ten clusters of six words each for each of the five bands described above. Each band is thus represented by 60 words. These words involve 30 nouns, 18 verbs, and 12 adjectives and are chosen at random from the 1,000 words of a band. Each cluster focuses on one part of speech, e. g., noun.

Each cluster contains six words and three synonyms, paraphrases, or gapped sentences (targets). Three of the six words are keys, i. e., they correspond to the three targets, while three words are additional distractors. For each target, the same six words are presented as multiple-choice options, one of which needs to be selected for each target. Each band, accordingly, consists of 30 items (targets). The maximum score per band is 30, i. e., 3 points per cluster. The maximum composite score for all five bands is 150, i. e., five times 30.

The productive vocabulary size test consists of 18 sentences, one for each targeted word, which is partially missing. Each band is thus represented by 18 words. These words involve 9

nouns, 6 verbs, and 3 adjectives chosen at random from the 1,000 words of a band. The maximum score per band is 18. The maximum composite score for all five bands is 90, i. e., five times 18.

The targeted words appear towards the end of the sentence to establish their meaning. The first few letters are given to disambiguate the word from other possible words. As many letters are provided as needed to disambiguate any given word, up to 50 % of the letters of the word, i. e., if a word consists of an odd number of letters, a maximum of half of the letters minus one are provided. All words of a particular sentence are part of the same band or a more frequent band. In Bands 1 and 2, the partially missing words are uninflected: verbs, for example, appear in their infinitive form. In Bands 3, 4, and 5, partially missing words may be inflected, i. e., grammatical knowledge may be required. Words are scored correct only if they are 100 % correct, including orthographic and grammatical correctness.

When test takers take the VST, their results are stored anonymously, without collecting any personal or technical information, to improve the VST. In the present report, data collected between April 2019 and March 2021 were analyzed to examine the overall validity and reliability of the VST and to identify poorly performing items to be revised in a number of languages.

The maximum time allowed for each test (receptive or productive) is 30 minutes. Tests that were completed in less

than five minutes were removed to reduce the number of test takers who were responding only to a few items. To be included in this report, each receptive or productive test needed to have had at least 30 results. A total of eight receptive tests complied with this requirement, i. e., Arabic, Chinese, English, French, German, Italian, Russian, and Spanish as well as a total of four productive tests, i. e., English, French, German, and Spanish.

Arabic: Receptive Vocabulary Size Test 1

The Arabic 1 Vocabulary Size Test (VST) 1 was modeled after the English Vocabulary Levels Test pioneered by Paul Nation (Nation, 1990). It uses the word frequency list of the Routledge Frequency Dictionary of Arabic (Buckwalter & Parkinson, 2011), which contains the most frequent 5,000 words of Arabic, and measures how many of them are known. In this chapter, evidence of validity and reliability of the receptive Arabic 1 VST is presented. No data were available for the productive Arabic 1 VST.

The receptive Arabic 1 VST consists of five bands: the most frequent 1,000; 1,001 to 2,000; 2,001 to 3,000; 3,001 to 4,000; and 4,001 to 5,000 words. It includes ten clusters of six words each for each of the five bands. Each band is thus represented by 60 words. These words involve 30 nouns, 18 verbs, and 12 adjectives and are chosen at random from the 1,000 words of a band. Each cluster focuses on one part of speech, e. g., noun.

Each cluster contains six words and three synonyms, paraphrases, or gapped sentences (targets). Three of the six words are keys, i. e., they correspond to the three targets, while three words are additional distractors. For each target, the same six words are presented as multiple-choice options, one of which needs to be selected for each target. Each band, accordingly, consists of 30 items (targets). The maximum score per band is 30, i. e., 3 points per cluster. The maximum composite score for all five bands is 150, i. e., five times 30.

When test takers take the VST, their results are stored anonymously, without collecting any personal or technical information, to improve the VST. In the present report, data collected between April 2019 and March 2021 were analyzed to examine the overall validity and reliability of the receptive Arabic 1 VST and to identify poorly performing items to be revised.

Correct responses were coded as 1 and incorrect responses as 0. Items that were not attempted were left blank. The maximum time allowed for the five-band test is 30 minutes. Tests that were completed in less than five minutes were removed to reduce the number of test takers who were responding only to a few items. Table A-1 shows the descriptive statistics of the receptive Arabic 1 VST.

Table A-1: Descriptive Statistics of the Receptive Arabic 1 Vocabulary Size Test

N	Mean	SE of Mean	Median	Std. Dev.	Minimum	Maximum
90	103.24	4.98	125.50	47.28	0	148

Table A-1 shows that there were 90 test takers. Total scores ranged from 0 to 148, covering almost the complete breadth of scores. The mean and the median were within the upper third of the score range, indicating that there was a large number of test takers with high receptive vocabulary sizes.

To examine the overall reliability of the receptive Arabic 1 VST, Cronbach's alpha between the five bands of the receptive

test was calculated. Cronbach's alpha is a measure of consistency, i. e., how consistent the results of all bands are to each other. It is commonly used as a measure of interrater reliability. Because it is a measure of internal consistency, it may be considered a measure of (internal) validity, i. e., it assesses how well different item sets measure the construct. If alpha is high, it may be assumed that all items measure the same construct, in this case receptive vocabulary size. Cronbach's alpha above 0.7 is considered acceptable, above 0.8, it is considered to be good, and above 0.9 very good. Table A-2 shows the number of tests administered, Cronbach's alpha, and the number of items, in this case, bands.

Table A-2: Cronbach's Alpha as a Measure of the Validity and Reliability of the Receptive Arabic 1 Vocabulary Size Test

N of Tests	Cronbach's Alpha	N of Items
67	0.96	5

Table A-2 shows that the reliability and internal validity of the receptive Arabic 1 VST was above 0.9 (very good), which supports the claim that it is highly valid and reliable. Note that the number of tests is different from the number of tests in Table A-1, because to calculate alpha across all bands, all bands need to have values. Some test takers only attempted one or more bands but not all five. For these test takers, Cronbach's alpha of all five bands could not be calculated.

To examine the internal consistency of each band, Cronbach's alpha was calculated for each band of 1,000 words.

Each band consists of 30 items. Table A-3 shows the number of test takers, Cronbach's alpha, and the number of items for each band of the receptive Arabic 1 VST.

Table A-3: Cronbach's Alpha for Each Band of the Receptive Arabic 1 Vocabulary Size Test

Band	N of Tests	Alpha	N of Items
1	66	0.93	30
2	60	0.94	30
3	62	0.96	30
4	59	0.96	30
5	51	0.98	30
Total	90		

Table A-3 shows that the internal consistency of each band was above 0.9 (very good), for some bands considerably above 0.9. To examine the goodness-of-fit of each individual item, the following statistics on the relationship between each individual item and all items of a band were calculated: the scale mean if the item was deleted; the corrected item-total correlation; and Cronbach's alpha if the item was deleted. Items below 0.3 in the column ***Corrected Item-Total Correlation*** do not correlate well with the overall score and, therefore, provide cause for concern (Field, 2018: 605). Items above the overall Cronbach's alpha of each band in the column ***Cronbach's Alpha if Item Deleted*** are also problematic, because if their removal raises alpha, then they

are less reliable than the average item (Field, 2018: 605). Tables A-4 to A-8 show the summary item statistics for each band of the receptive Arabic 1 VST. Cells with misfitting values are set in bold and red.

Table A-4: Summary Item Statistics for the Receptive Arabic 1 VST Band 1

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@1a	25.27	0.61	0.93
@1b	25.27	0.68	0.93
@1c	25.29	0.67	0.93
@2a	25.32	0.53	0.93
@2b	25.35	0.57	0.93
@2c	25.33	0.50	0.93
@3a	25.24	0.60	0.93
@3b	25.29	0.68	0.93
@3c	25.88	**0.15**	**0.94**
@4a	25.38	0.53	0.93
@4b	25.45	0.46	0.93
@4c	25.39	0.61	0.93
@5a	25.35	0.48	0.93

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@5b	25.47	0.55	0.93
@5c	25.32	0.72	0.93
@6a	25.45	0.38	0.93
@6b	25.32	0.63	0.93
@6c	25.27	0.67	0.93
@7a	25.36	0.72	0.92
@7b	25.38	0.61	0.93
@7c	25.35	0.58	0.93
@8a	25.33	0.66	0.93
@8b	25.29	0.63	0.93
@8c	25.27	0.58	0.93
@9a	25.29	0.35	0.93
@9b	25.33	0.70	0.93
@9c	25.26	0.76	0.93
@10a	25.32	0.71	0.93
@10b	25.29	0.38	0.93
@10c	25.47	0.47	0.93

Table A-5: Summary Item Statistics for the Receptive Arabic 1 VST Band 2

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@11a	24.85	0.76	0.94
@11b	24.90	0.60	0.94
@11c	24.92	0.82	0.93
@12a	24.87	0.77	0.93
@12b	24.85	0.76	0.94
@12c	24.87	0.50	0.94
@13a	24.97	0.45	0.94
@13b	24.88	0.77	0.93
@13c	25.05	0.41	0.94
@14a	24.88	0.71	0.93
@14b	24.92	0.74	0.93
@14c	25.45	**0.23**	0.94
@15a	25.05	0.58	0.94
@15b	24.93	0.73	0.93
@15c	25.02	0.51	0.94
@16a	24.85	0.76	0.94

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@16b	25.05	0.42	0.94
@16c	25.30	**0.13**	0.94
@17a	24.88	0.66	0.94
@17b	25.15	0.35	0.94
@17c	24.90	0.78	0.93
@18a	24.90	0.56	0.94
@18b	24.88	0.78	0.93
@18c	24.90	0.63	0.94
@19a	24.92	0.80	0.93
@19b	24.93	0.76	0.93
@19c	24.87	0.40	0.94
@20a	24.90	0.69	0.93
@20b	24.93	0.76	0.93
@20c	24.92	0.80	0.93

Table A-6: Summary Item Statistics for the Receptive Arabic 1 VST Band 3

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@21a	25.76	0.62	0.96

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@21b	25.81	0.69	0.96
@21c	25.79	0.73	0.96
@22a	25.77	0.81	0.96
@22b	25.77	0.68	0.96
@22c	25.82	0.75	0.96
@23a	25.73	0.81	0.96
@23b	26.13	**0.25**	0.96
@23c	25.73	0.62	0.96
@24a	25.71	0.80	0.96
@24b	25.74	0.69	0.96
@24c	25.69	0.67	0.96
@25a	25.71	0.71	0.96
@25b	25.71	0.71	0.96
@25c	25.71	0.80	0.96
@26a	25.84	0.62	0.96
@26b	25.81	0.68	0.96
@26c	25.77	0.67	0.96
@27a	25.77	0.76	0.96

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@27b	25.77	0.75	0.96
@27c	25.73	**0.28**	0.96
@28a	25.74	0.78	0.96
@28b	25.71	0.80	0.96
@28c	25.76	0.63	0.96
@29a	25.84	0.74	0.96
@29b	25.90	0.51	0.96
@29c	25.76	0.84	0.96
@30a	25.71	0.69	0.96
@30b	25.77	0.71	0.96
@30c	25.71	0.70	0.96

Table A-7: Summary Item Statistics for the Receptive Arabic 1 VST Band 4

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@31a	24.47	0.80	0.96
@31b	24.53	0.83	0.96
@31c	24.51	0.89	0.96
@32a	24.54	0.79	0.96

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@32b	24.53	0.74	0.96
@32c	24.54	0.46	0.96
@33a	24.54	0.68	0.96
@33b	24.58	0.84	0.96
@33c	24.49	0.42	0.96
@34a	24.51	0.82	0.96
@34b	24.46	0.74	0.96
@34c	24.49	0.82	0.96
@35a	24.54	0.63	0.96
@35b	24.49	0.70	0.96
@35c	24.54	0.55	0.96
@36a	24.54	0.57	0.96
@36b	24.49	0.89	0.96
@36c	24.49	0.89	0.96
@37a	24.81	0.30	0.96
@37b	24.54	0.83	0.96
@37c	24.53	0.77	0.96
@38a	24.78	0.46	0.96

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@38b	24.78	0.53	0.96
@38c	24.54	0.81	0.96
@39a	24.64	0.58	0.96
@39b	24.71	0.45	0.96
@39c	24.61	0.73	0.96
@40a	24.47	0.68	0.96
@40b	24.59	0.55	0.96
@40c	24.49	0.89	0.96

Table A-8: Summary Item Statistics for the Receptive Arabic 1 VST Band 5

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@41a	25.39	0.67	0.98
@41b	25.47	0.66	0.98
@41c	25.31	0.83	0.98
@42a	25.45	0.52	0.98
@42b	25.41	0.78	0.98
@42c	25.41	0.71	0.98
@43a	25.31	0.89	0.98

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@43b	25.53	0.56	0.98
@43c	25.33	0.78	0.98
@44a	25.35	0.91	0.98
@44b	25.33	0.96	0.98
@44c	25.39	0.73	0.98
@45a	25.39	0.76	0.98
@45b	25.31	0.88	0.98
@45c	25.35	0.88	0.98
@46a	25.31	0.88	0.98
@46b	25.33	0.80	0.98
@46c	25.31	0.89	0.98
@47a	25.31	0.88	0.98
@47b	25.33	0.67	0.98
@47c	25.37	0.68	0.98
@48a	25.29	0.82	0.98
@48b	25.29	0.82	0.98
@48c	25.31	0.89	0.98
@49a	25.33	0.96	0.98

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@49b	25.33	0.96	0.98
@49c	25.33	0.96	0.98
@50a	25.49	0.59	0.98
@50b	25.33	0.96	0.98
@50c	25.35	0.52	0.98

Tables A-4 to A-8 show that there were only five items in the receptive Arabic 1 Vocabulary Size Test with conspicuous values. They will be inspected and revised, if necessary, for version 2 of the receptive Arabic 1 VST. Five items out of 150, however, do not threaten the validity or reliability of the present version of the receptive Arabic 1 VST. The data analyzed in this chapter, therefore, provide strong evidence that the present version (Arabic 1.1) of the receptive Arabic 1 VST exhibits excellent psychometric properties, meeting all professional standards of reliability and validity required in high-stakes testing.

Chinese: Receptive Vocabulary Size Test 1

The Chinese 1 Vocabulary Size Test (VST) was modeled after the English Vocabulary Levels Test pioneered by Paul Nation (Nation, 1990). It uses the word frequency list of the Routledge Frequency Dictionary of Chinese (Xiao, Rayson, & McEnery, 2009), which contains the most frequent 5,000 words of Chinese, and measures how many of them are known. In this chapter, evidence of validity and reliability of the receptive Chinese 1 VST is presented. No data were available for the productive Chinese 1 VST.

The receptive Chinese 1 VST consists of five bands: the most frequent 1,000; 1,001 to 2,000; 2,001 to 3,000; 3,001 to 4,000; and 4,001 to 5,000 words. It includes ten clusters of six words each for each of the five bands. Each band is thus represented by 60 words. These words involve 30 nouns, 18 verbs, and 12 adjectives and are chosen at random from the 1,000 words of a band. Each cluster focuses on one part of speech, e. g., noun.

Each cluster contains six words and three synonyms, paraphrases, or gapped sentences (targets). Three of the six words are keys, i. e., they correspond to the three targets, while three words are additional distractors. For each target, the same six words are presented as multiple-choice options, one of which needs to be selected for each target. Each band, accordingly, consists of 30 items (targets). The maximum score per band is 30, i. e., 3 points per cluster. The maximum composite score for all five bands is 150, i. e., five times 30.

When test takers take the VST, their results are stored anonymously, without collecting any personal or technical information, to improve the VST. In the present report, data collected between April 2019 and March 2021 were analyzed to examine the overall validity and reliability of the receptive Chinese 1 VST and to identify poorly performing items to be revised.

Correct responses were coded as 1 and incorrect responses as 0. Items that were not attempted were left blank. The maximum time allowed for the five-band test is 30 minutes. Tests that were completed in less than five minutes were removed to reduce the number of test takers who were responding only to a few items. Table C-1 shows the descriptive statistics of the receptive Chinese 1 VST.

Table C-1: Descriptive Statistics of the Receptive Chinese 1 Vocabulary Size Test

N	Mean	SE of Mean	Median	Std. Dev.	Minimum	Maximum
72	92.10	6.61	101.50	56.04	0	150

Table C-1 shows that there were 72 test takers. Total scores ranged from 0 to 150, covering the complete breadth of scores. The mean and the median were close to or within the upper third of the score range, indicating that there was a large number of test takers with high receptive vocabulary sizes.

To examine the overall reliability of the receptive Chinese 1 VST, Cronbach's alpha between the five bands of the receptive

test was calculated. Cronbach's alpha is a measure of consistency, i. e., how consistent the results of all bands are to each other. It is commonly used as a measure of interrater reliability. Because it is a measure of internal consistency, it may be considered a measure of (internal) validity, i. e., it assesses how well different item sets measure the construct. If alpha is high, it may be assumed that all items measure the same construct, in this case receptive vocabulary size. Cronbach's alpha above 0.7 is considered acceptable, above 0.8, it is considered to be good, and above 0.9 very good. Table C-2 shows the number of tests administered, Cronbach's alpha, and the number of items, in this case, bands.

Table C-2: Cronbach's Alpha as a Measure of the Validity and Reliability of the Receptive Chinese 1 Vocabulary Size Test

N of Tests	Cronbach's Alpha	N of Items
46	0.93	5

Table C-2 shows that the reliability and internal validity of the receptive Chinese 1 VST was above 0.9 (very good), which supports the claim that it is highly valid and reliable. Note that the number of tests is different from the number of tests in Table C-1, because to calculate alpha across all bands, all bands need to have values. Some test takers only attempted one or more bands but not all five. For these test takers, Cronbach's alpha of all five bands could not be calculated.

To examine the internal consistency of each band, Cronbach's alpha was calculated for each band of 1,000 words.

Each band consists of 30 items. Table C-3 shows the number of test takers, Cronbach's alpha, and the number of items for each band of the receptive Chinese 1 VST.

Table C-3: Cronbach's Alpha for Each Band of the Receptive Chinese 1 Vocabulary Size Test

Band	N of Tests	Alpha	N of Items
1	55	0.96	30
2	47	0.97	30
3	45	0.98	30
4	39	0.98	30
5	37	0.98	30
Total	72		

Table C-3 shows that the internal consistency of each band was considerably above 0.9, i. e., very good, for all five bands. To examine the goodness-of-fit of each individual item, the following statistics on the relationship between each individual item and all items of a band were calculated: the scale mean if the item was deleted; the corrected item-total correlation; and Cronbach's alpha if the item was deleted. Items below 0.3 in the column ***Corrected Item-Total Correlation*** do not correlate well with the overall score and, therefore, provide cause for concern (Field, 2018: 605). Items above the overall Cronbach's alpha of each band in the column ***Cronbach's Alpha if Item Deleted*** are also problematic, because if their removal raises alpha, then they

are less reliable than the average item (Field, 2018: 605). Tables C-4 to C-8 show the summary item statistics for each band of the receptive Chinese 1 VST. Cells with misfitting values are set in bold and red.

Table C-4: Summary Item Statistics for the Receptive Chinese 1 VST Band 1

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@1a	23.16	0.57	0.96
@1b	23.09	0.84	0.95
@1c	23.16	0.73	0.95
@2a	23.45	0.43	0.96
@2b	23.20	0.59	0.96
@2c	22.91	**0.29**	0.96
@3a	22.98	0.69	0.96
@3b	23.09	0.70	0.95
@3c	23.07	0.77	0.95
@4a	23.07	0.79	0.95
@4b	22.96	0.57	0.96
@4c	23.05	0.67	0.95
@5a	23.04	0.80	0.95

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@5b	22.96	0.49	0.96
@5c	23.05	0.76	0.95
@6a	23.13	0.74	0.95
@6b	23.05	0.66	0.96
@6c	23.07	0.82	0.95
@7a	23.04	0.66	0.96
@7b	23.16	0.62	0.96
@7c	23.13	0.49	0.96
@8a	23.04	0.62	0.96
@8b	23.13	0.85	0.95
@8c	23.00	0.72	0.95
@9a	23.11	0.73	0.95
@9b	23.36	0.47	0.96
@9c	23.40	0.49	0.96
@10a	22.96	0.61	0.96
@10b	22.95	0.58	0.96
@10c	23.04	0.67	0.95

Table C-5: Summary Item Statistics for the Receptive Chinese 1 VST Band 2

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@11a	25.85	0.58	0.97
@11b	25.89	0.62	0.97
@11c	25.89	0.93	0.96
@12a	25.91	0.63	0.97
@12b	25.85	0.81	0.97
@12c	25.91	0.91	0.97
@13a	25.89	0.93	0.96
@13b	25.85	0.78	0.97
@13c	25.94	0.85	0.97
@14a	25.89	0.82	0.97
@14b	25.91	0.45	0.97
@14c	25.81	0.63	0.97
@15a	25.79	0.57	0.97
@15b	25.81	0.63	0.97
@15c	25.83	0.72	0.97
@16a	25.85	0.83	0.97
@16b	25.85	0.78	0.97

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@16c	25.79	0.57	0.97
@17a	25.89	0.93	0.96
@17b	25.85	0.78	0.97
@17c	25.85	0.62	0.97
@18a	25.89	0.82	0.97
@18b	25.83	0.75	0.97
@18c	25.94	0.87	0.97
@19a	25.98	0.65	0.97
@19b	25.94	0.64	0.97
@19c	26.04	0.34	0.97
@20a	25.83	0.70	0.97
@20b	25.85	0.65	0.97
@20c	25.79	0.57	0.97

Table C-6: Summary Item Statistics for the Receptive Chinese 1 VST Band 3

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@21a	26.53	0.50	0.98
@21b	26.53	0.86	0.97

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@21c	26.47	0.71	0.98
@22a	26.47	0.40	0.98
@22b	26.47	0.70	0.98
@22c	26.51	0.97	0.97
@23a	26.47	0.59	0.98
@23b	26.49	0.90	0.97
@23c	26.49	0.90	0.97
@24a	26.47	0.71	0.98
@24b	26.51	0.97	0.97
@24c	26.47	0.73	0.98
@25a	26.49	0.90	0.97
@25b	26.51	0.97	0.97
@25c	26.44	0.52	0.98
@26a	26.49	0.90	0.97
@26b	26.49	0.90	0.97
@26c	26.56	0.78	0.98
@27a	26.51	0.97	0.97
@27b	26.56	0.80	0.97

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@27c	26.47	0.61	0.98
@28a	26.56	0.80	0.97
@28b	26.47	0.70	0.98
@28c	26.56	0.57	0.98
@29a	26.67	0.61	0.98
@29b	26.58	0.77	0.98
@29c	26.56	0.62	0.98
@30a	26.51	0.97	0.97
@30b	26.49	0.81	0.97
@30c	26.49	0.78	0.98

Table C-7: Summary Item Statistics for the Receptive Chinese 1 VST Band 4

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@31a	26.49	0.87	0.97
@31b	26.46	0.60	0.98
@31c	26.49	0.86	0.97
@32a	26.49	0.87	0.97
@32b	26.44	0.58	0.98

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@32c	26.46	0.71	0.98
@33a	26.49	0.82	0.97
@33b	26.51	0.98	0.97
@33c	26.54	0.87	0.97
@34a	26.54	0.87	0.97
@34b	26.49	0.53	0.98
@34c	26.44	0.37	0.98
@35a	26.51	0.98	0.97
@35b	26.49	0.86	0.97
@35c	26.49	0.87	0.97
@36a	26.51	0.73	0.98
@36b	26.54	0.67	0.98
@36c	26.67	**0.25**	0.98
@37a	26.44	0.58	0.98
@37b	26.51	0.98	0.97
@37c	26.49	0.82	0.97
@38a	26.49	0.47	0.98
@38b	26.46	0.75	0.98

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@38c	26.49	0.74	0.97
@39a	26.51	0.98	0.97
@39b	26.46	0.71	0.98
@39c	26.46	0.75	0.98
@40a	26.49	0.86	0.97
@40b	26.51	0.98	0.97
@40c	26.56	0.80	0.97

Table C-8: Summary Item Statistics for the Receptive Chinese 1 VST Band 5

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@41a	26.22	0.54	0.98
@41b	26.08	0.81	0.98
@41c	26.08	0.78	0.98
@42a	26.05	0.72	0.98
@42b	26.11	0.98	0.98
@42c	26.11	0.98	0.98
@43a	26.11	0.68	0.98
@43b	26.11	0.98	0.98

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@43c	26.08	0.81	0.98
@44a	26.19	0.70	0.98
@44b	26.11	0.98	0.98
@44c	26.03	0.52	0.98
@45a	26.11	0.98	0.98
@45b	26.11	0.81	0.98
@45c	26.08	0.89	0.98
@46a	26.11	0.71	0.98
@46b	26.11	0.98	0.98
@46c	26.11	0.81	0.98
@47a	26.11	0.74	0.98
@47b	26.11	0.98	0.98
@47c	26.14	0.67	0.98
@48a	26.11	0.98	0.98
@48b	26.11	0.72	0.98
@48c	26.05	0.60	0.98
@49a	26.11	0.69	0.98
@49b	26.08	0.89	0.98

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@49c	26.08	0.89	0.98
@50a	26.08	0.54	0.98
@50b	26.08	0.89	0.98
@50c	26.05	0.72	0.98

Tables C-4 to C-8 show that there were only two items in the receptive Chinese 1 Vocabulary Size Test with conspicuous values. They will be inspected and revised, if needed, for version 2 of the receptive Chinese 1 VST. Two items out of 150, however, do not threaten the validity or reliability of the present version of the receptive Chinese 1 VST. The data analyzed in this chapter, therefore, provide strong evidence that the present version (Chinese 1.1) of the receptive Chinese 1 VST exhibits excellent psychometric properties, meeting all professional standards of reliability and validity required in high-stakes testing.

English: Receptive and Productive Vocabulary Size Tests 1

The English 1 Vocabulary Size Test (VST) was modeled after the English Vocabulary Levels Test pioneered by Paul Nation (Nation, 1990). It uses the word frequency list of Leech, Rayson, & Wilson (2001), which is based on the British National Corpus, containing the most frequent 5,000 words of English. The VST measures how many of them are known. In this chapter, evidence of validity and reliability of both, the receptive and productive English 1 VST is presented. When test takers take the VST, their results are stored anonymously, without collecting any personal or technical information, to improve the VST. In the present report, data collected between April 2019 and March 2021 were analyzed to examine the overall validity and reliability of the receptive and productive English 1 VST and to identify poorly performing items to be revised.

The Receptive English 1 Vocabulary Size Test

The receptive English 1 VST consists of five bands: the most frequent 1,000; 1,001 to 2,000; 2,001 to 3,000; 3,001 to 4,000; and 4,001 to 5,000 words. It includes ten clusters of six words each for each of the five bands. Each band is thus represented by 60 words. These words involve 30 nouns, 18 verbs, and 12 adjectives and are chosen at random from the 1,000 words of a band. Each cluster focuses on one part of speech, e. g., noun.

Each cluster contains six words and three synonyms, paraphrases, or gapped sentences (targets). Three of the six words are keys, i. e., they correspond to the three targets, while three words are additional distractors. For each target, the same six words are presented as multiple-choice options, one of which needs to be selected for each target. Each band, accordingly, consists of 30 items (targets). The maximum score per band is 30, i. e., 3 points per cluster. The maximum composite score for all five bands is 150, i. e., five times 30.

Correct responses were coded as 1 and incorrect responses as 0. Items that were not attempted were left blank. The maximum time allowed for the five-band test is 30 minutes. Tests that were completed in less than five minutes were removed to reduce the number of test takers who were responding only to a few items. Table E-1 shows the descriptive statistics of the receptive English 1 VST.

Table E-1: Descriptive Statistics of the Receptive English 1 Vocabulary Size Test

N	Mean	SE of Mean	Median	Std. Dev.	Minimum	Maximum
1346	104.99	1.21	125	44.39	0	150

Table E-1 shows that there were 1346 test takers. Total scores ranged from 0 to 150, covering the complete breadth of scores. The mean and the median were within the upper third of the score range with a standard deviation of 44, indicating that there was a large number of test takers with high receptive vocabulary sizes.

To examine the overall reliability of the receptive English 1 VST, Cronbach's alpha between the five bands of the receptive test was calculated. Cronbach's alpha is a measure of consistency, i. e., how consistent the results of all bands are to each other. It is commonly used as a measure of interrater reliability. Because it is a measure of internal consistency, it may be considered a measure of (internal) validity, i. e., it assesses how well different item sets measure the construct. If alpha is high, it may be assumed that all items measure the same construct, in this case receptive vocabulary size. Cronbach's alpha above 0.7 is considered acceptable, above 0.8, it is considered to be good, and above 0.9 very good. Table E-2 shows the number of tests administered, Cronbach's alpha, and the number of items, in this case, bands.

Table E-2: Cronbach's Alpha as a Measure of the Validity and Reliability of the Receptive English 1 Vocabulary Size Test

N of Tests	Cronbach's Alpha	N of Items
1112	0.94	5

Table E-2 shows that the reliability and internal validity of the receptive English 1 VST is above 0.9 (very good), which supports the claim that it is highly valid and reliable. Note that the number of tests is different from the number of tests in Table E-1, because to calculate alpha across all bands, all bands need to have values. Some test takers only attempted one or more bands but not all five. For these test takers, Cronbach's alpha of all five bands could not be calculated.

To examine the internal consistency of each band, Cronbach's alpha was calculated for each band of 1,000 words. Each band consists of 30 items. Table E-3 shows the number of test takers, Cronbach's alpha, and the number of items for each band of the receptive English 1 VST.

Table E-3: Cronbach's Alpha for Each Band of the Receptive English 1 Vocabulary Size Test

Band	N of Tests	Alpha	N of Items
1	54	0.84	30
2	51	0.89	30
3	988	0.89	30
4	971	0.90	30
5	876	0.92	30
Total	1346		

Table E-3 shows that the internal consistency of each band was above 0.8 for Band 1 (good) and close to or above 0.9 (very good) for all other bands. To examine the goodness-of-fit of each individual item, the following statistics on the relationship between each individual item and all items of a band were calculated: the scale mean if the item was deleted; the corrected item-total correlation; and Cronbach's alpha if the item was deleted. Items below 0.3 in the column ***Corrected Item-Total Correlation*** do not correlate well with the overall score and, therefore, provide cause for concern (Field, 2018: 605). Items

above the overall Cronbach's alpha of each band in the column ***Cronbach's Alpha if Item Deleted*** are also problematic, because if their removal raises alpha, then they are less reliable than the average item (Field, 2018: 605). Tables E-4 to E-8 show the summary item statistics for each band of the receptive English 1 VST. Cells with misfitting values are set in bold and red.

Table E-4: Summary Item Statistics for the Receptive English 1 VST Band 1

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@1a	27.44	0.57	0.82
@1b	27.37	**0.00**	0.84
@1c	27.46	0.50	0.83
@2a	27.37	**0.00**	0.84
@2b	27.37	**0.00**	0.84
@2c	27.43	**0.04**	0.84
@3a	27.37	**0.00**	0.84
@3b	27.37	**0.00**	0.84
@3c	27.41	0.58	0.82
@4a	27.39	0.73	0.82
@4b	27.41	**0.11**	0.84
@4c	27.37	**0.00**	0.84

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@5a	27.39	0.73	0.82
@5b	27.70	0.49	0.83
@5c	27.57	0.46	0.83
@6a	27.39	0.73	0.82
@6b	27.41	0.62	0.82
@6c	27.44	0.57	0.82
@7a	27.39	**-0.08**	0.84
@7b	27.41	**0.18**	0.84
@7c	27.39	0.73	0.82
@8a	27.41	**0.11**	0.84
@8b	27.69	0.44	0.83
@8c	27.39	0.73	0.82
@9a	27.37	**0.00**	0.84
@9b	27.37	**0.00**	0.84
@9c	27.43	0.49	0.83
@10a	27.46	0.64	0.82
@10b	27.37	**0.00**	0.84
@10c	27.41	0.62	0.82

Table E-5: Summary Item Statistics for the Receptive English 1 VST Band 2

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@11a	26.69	**-0.01**	0.89
@11b	26.69	**0.06**	0.89
@11c	26.67	**0.00**	0.89
@12a	26.73	**0.16**	0.89
@12b	26.71	0.50	0.89
@12c	26.82	0.65	0.88
@13a	26.82	0.61	0.88
@13b	26.69	**0.21**	0.89
@13c	26.75	0.50	0.89
@14a	26.69	0.72	0.89
@14b	26.94	0.58	0.89
@14c	26.75	0.48	0.89
@15a	26.71	0.62	0.89
@15b	26.71	0.67	0.88
@15c	26.69	0.72	0.89
@16a	26.69	0.72	0.89
@16b	26.73	0.53	0.89

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@16c	26.67	**0.00**	0.89
@17a	26.78	0.61	0.88
@17b	26.73	0.69	0.88
@17c	26.71	0.62	0.89
@18a	26.67	**0.00**	0.89
@18b	26.73	0.43	0.89
@18c	26.69	0.72	0.89
@19a	26.80	**0.23**	0.89
@19b	26.82	0.36	0.89
@19c	26.78	0.38	0.89
@20a	26.82	0.44	0.89
@20b	26.78	0.61	0.88
@20c	26.92	0.55	0.89

Table E-6: Summary Item Statistics for the Receptive English 1 VST Band 3

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@21a	23.80	0.49	0.88
@21b	23.74	0.45	0.89

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@21c	23.75	0.45	0.89
@22a	23.95	0.53	0.88
@22b	23.85	0.57	0.88
@22c	23.73	0.40	0.89
@23a	23.88	0.57	0.88
@23b	23.76	0.46	0.89
@23c	23.87	0.62	0.88
@24a	23.75	0.52	0.89
@24b	23.74	0.41	0.89
@24c	24.38	**0.27**	0.89
@25a	23.79	0.51	0.88
@25b	23.77	0.54	0.88
@25c	23.77	0.51	0.88
@26a	23.76	0.41	0.89
@26b	24.35	0.35	0.89
@26c	24.14	0.39	0.89
@27a	24.29	0.30	0.89
@27b	23.79	0.56	0.88

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@27c	23.79	0.46	0.89
@28a	23.78	0.44	0.89
@28b	23.80	0.59	0.88
@28c	23.76	0.48	0.89
@29a	23.74	0.45	0.89
@29b	23.91	0.57	0.88
@29c	24.04	0.39	0.89
@30a	23.79	0.54	0.88
@30b	24.27	0.35	0.89
@30c	23.77	0.47	0.89

Table E-7: Summary Item Statistics for the Receptive English 1 VST Band 4

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@31a	25.51	0.53	0.90
@31b	25.47	**0.26**	0.90
@31c	25.57	0.42	0.90
@32a	25.57	0.50	0.90
@32b	25.54	0.46	0.90

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@32c	25.47	0.52	0.90
@33a	25.47	0.43	0.90
@33b	25.60	0.43	0.90
@33c	25.56	0.53	0.90
@34a	25.68	0.50	0.90
@34b	25.47	0.44	0.90
@34c	25.49	0.42	0.90
@35a	25.56	0.53	0.90
@35b	25.57	0.53	0.90
@35c	25.52	0.54	0.90
@36a	25.55	0.54	0.90
@36b	25.64	0.42	0.90
@36c	25.63	0.52	0.90
@37a	25.53	0.45	0.90
@37b	25.52	0.57	0.90
@37c	25.60	0.37	0.90
@38a	25.53	0.55	0.90
@38b	25.57	0.54	0.90

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@38c	25.61	0.55	0.90
@39a	25.70	0.51	0.90
@39b	25.64	0.49	0.90
@39c	25.52	**0.24**	0.90
@40a	25.61	0.39	0.90
@40b	25.63	0.36	0.90
@40c	25.61	0.57	0.90

Table E-8: Summary Item Statistics for the Receptive English 1 VST Band 5

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@41a	25.84	0.58	0.92
@41b	25.95	0.51	0.92
@41c	25.96	0.48	0.92
@42a	25.75	0.63	0.92
@42b	25.74	0.57	0.92
@42c	25.72	0.54	0.92
@43a	25.71	0.48	0.92
@43b	25.78	0.55	0.92

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@43c	25.78	0.38	0.92
@44a	25.74	0.40	0.92
@44b	25.80	0.51	0.92
@44c	25.76	**0.27**	0.92
@45a	25.74	0.56	0.92
@45b	25.74	0.57	0.92
@45c	25.74	0.48	0.92
@46a	25.88	0.52	0.92
@46b	25.79	0.55	0.92
@46c	25.88	0.62	0.92
@47a	25.80	0.61	0.92
@47b	25.86	0.64	0.92
@47c	25.82	0.49	0.92
@48a	26.02	0.54	0.92
@48b	25.75	0.51	0.92
@48c	25.79	0.55	0.92
@49a	25.76	0.62	0.92
@49b	25.73	0.53	0.92

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@49c	25.74	0.59	0.92
@50a	25.85	0.44	0.92
@50b	25.76	0.55	0.92
@50c	25.75	0.52	0.92

Tables E-4 to E-8 show that there were 22 items in Bands 1 and 2 and 4 items in Bands 3-5 with conspicuous values in the receptive English 1 Vocabulary Size Test. Note the relatively low number of tests in Bands 1 and 2 compared with the other three bands. These items were inspected. None of these items showed any obvious issues. The data analyzed in this section and the subsequent inspection of flagged items provide strong evidence that the receptive English 1 VST exhibits excellent psychometric properties, meeting all professional standards of reliability and validity required in high-stakes testing.

The Productive English 1 Vocabulary Size Test

The productive vocabulary size test consists of 18 sentences, one for each targeted word, which is partially missing. Each band is thus represented by 18 words. These words involve 9 nouns, 6 verbs, and 3 adjectives chosen at random from the 1,000 words of a band. The maximum score per band is 18. The maximum composite score for all five bands is 90, i. e., five times 18.

The targeted words appear towards the end of the sentence to establish their meaning. The first few letters are given to disambiguate the word from other possible words. As many letters are provided as needed to disambiguate any given word, up to 50 % of the letters of the word, i. e., if a word consists of an odd number of letters, a maximum of half of the letters minus one are provided. All words of a particular sentence are part of the same band or a more frequent band. In Bands 1 and 2, the partially missing words are uninflected: verbs, for example, appear in their infinitive form. In Bands 3, 4, and 5, partially missing words may be inflected, i. e., grammatical knowledge may be required. Words are scored correct only if they are 100 % correct, including orthographic and grammatical correctness.

Correct responses were coded as 1 and incorrect responses as 0. Items that were not attempted were left blank. Table E-9 shows the descriptive statistics of the productive English 1 VST. Tests that were completed in less than 5 minutes were removed.

Table E-9: Descriptive Statistics of the Productive English 1 Vocabulary Size Test

N	Mean	SE of Mean	Median	Std. Dev.	Minimum	Maximum
906	55.74	0.75	62	22.58	0	87

Table E-9 shows that there were 906 test takers. Total scores ranged from 0 to 87, all but covering the complete breadth of scores. The mean and the median were above the midpoint of

the scale (45) with a standard deviation of 23, indicating that there was a broad range of test takers.

To examine the overall reliability of the productive English 1 VST, Cronbach's alpha between the five bands of the productive test was calculated. Cronbach's alpha is a measure of consistency, i. e., how consistent the results of all bands are to each other. Because it is a measure of internal consistency, it may be considered a measure of (internal) validity, i. e., it assesses how well different item sets measure the construct. If alpha is high, it may be assumed that all items measure the same construct, in this case productive vocabulary size. Cronbach's alpha above 0.7 is considered acceptable, above 0.8, it is considered to be good, and above 0.9 very good. Table E-10 shows the number of tests administered, Cronbach's alpha, and the number of items, in this case bands.

Table E-10: Cronbach's Alpha as a Measure of the Validity and Reliability of the Productive English 1 Vocabulary Size Test

N of Tests	Cronbach's Alpha	N of Items
805	0.94	5

Table E-10 shows that the reliability and internal validity of the productive English 1 VST was above 0.9 (very good), which supports the claim that it is highly valid and reliable. Note that the number of tests is different from the number of tests in Table E-9, because to calculate alpha across all bands, all bands need to have values. Some test takers only attempted one or

more bands but not all five. For these test takers, Cronbach's alpha of all five bands could not be calculated.

To examine the internal consistency of each band, Cronbach's alpha was calculated for each band of 1,000 words. Each band consists of 18 items. Table E-11 shows the number of test takers, Cronbach's alpha, and the number of items for each band of the productive English 1 VST.

Table E-11: Cronbach's Alpha for Each Band of the Productive English 1 Vocabulary Size Test

Band	N of Tests	Alpha	N of Items
1	583	0.76	18
2	435	0.76	18
3	320	0.79	18
4	304	0.80	18
5	274	0.84	18
Total	906		

Table E-11 shows that the internal consistency of each band was mostly close to or above 0.8, i. e., close to or within the good range. Bands 1 and 2 were more on the acceptable (above 0.7) side.

To examine the goodness-of-fit of each individual item, the following statistics on the relationship between each individual item and all items of a band were calculated: the scale mean if

the item was deleted; the corrected item-total correlation; and Cronbach's alpha if the item was deleted. Items below 0.3 in the column ***Corrected Item-Total Correlation*** do not correlate well with the overall score and may, therefore, provide cause for concern (Field, 2018: 605). Items above the overall Cronbach's alpha of each band in the column ***Cronbach's Alpha if Item Deleted*** are also problematic, because if their removal raises alpha, then they are less reliable than the average item (Field, 2018: 605). Tables E-12 to E-16 show the summary item statistics for each band of the productive English 1 VST. Cells with misfitting values are set in bold and red.

Table E-12: Summary Item Statistics for the Productive English 1 VST Band 1

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@1	15.20	0.42	0.75
@2	15.25	0.32	0.75
@3	15.20	0.33	0.75
@4	15.32	0.31	0.76
@5	15.34	0.34	0.75
@6	15.28	0.35	0.75
@7	15.29	**0.23**	0.76
@8	15.18	0.53	0.74
@9	15.23	0.33	0.75

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@10	15.38	0.35	0.75
@11	15.16	0.61	0.75
@12	15.22	0.35	0.75
@13	15.33	**0.24**	0.76
@14	15.17	0.54	0.75
@15	15.22	0.31	0.75
@16	15.20	0.38	0.75
@17	15.21	0.43	0.75
@18	15.25	0.32	0.75

Table E-13: Summary Item Statistics for the Productive English 1 VST Band 2

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@19	14.52	0.39	0.75
@20	14.54	0.53	0.74
@21	14.57	0.40	0.74
@22	14.60	0.47	0.73
@23	14.59	0.34	0.75
@24	14.51	0.53	0.74

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@25	14.51	0.49	0.74
@26	14.57	0.32	0.75
@27	14.71	0.31	0.75
@28	14.68	0.33	0.75
@29	14.57	0.43	0.74
@30	14.85	**0.19**	**0.77**
@31	14.58	0.35	0.75
@32	15.17	**0.11**	**0.78**
@33	14.49	0.58	0.74
@34	14.53	0.37	0.75
@35	14.54	0.40	0.74
@36	14.58	0.35	0.75

Table E-14: Summary Item Statistics for the Productive English 1 VST Band 3

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@37	13.69	0.33	0.78
@38	13.84	0.40	0.78
@39	13.64	0.46	0.77

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@40	13.72	0.41	0.78
@41	13.64	0.49	0.77
@42	13.56	0.47	0.78
@43	13.51	0.36	0.78
@44	13.68	0.43	0.78
@45	13.66	0.39	0.78
@46	13.60	0.53	0.77
@47	13.52	0.40	0.78
@48	13.56	0.42	0.78
@49	13.76	**0.19**	**0.80**
@50	13.55	0.48	0.78
@51	13.52	0.51	0.78
@52	13.66	**0.27**	0.79
@53	13.56	0.44	0.78
@54	13.99	**0.06**	**0.81**

Table E-15: Summary Item Statistics for the Productive English 1 VST Band 4

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@55	13.52	0.33	0.79
@56	13.49	0.43	0.79
@57	13.57	0.32	0.79
@58	13.42	0.48	0.79
@59	13.50	0.38	0.79
@60	13.49	0.35	0.79
@61	13.75	0.33	0.79
@62	13.41	0.42	0.79
@63	13.54	0.31	0.79
@64	13.61	0.50	0.78
@65	13.71	0.43	0.79
@66	13.43	0.53	0.78
@67	13.88	0.35	0.79
@68	13.99	**0.28**	**0.80**
@69	13.43	0.53	0.78
@70	13.50	0.47	0.78
@71	13.42	0.53	0.79

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@72	13.78	0.45	0.78

Table E-16: Summary Item Statistics for the Productive English 1 VST Band 5

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@73	13.33	0.46	0.83
@74	13.27	0.48	0.83
@75	13.35	0.54	0.83
@76	13.29	0.52	0.83
@77	13.25	0.48	0.83
@78	13.73	**0.21**	**0.85**
@79	13.45	0.47	0.83
@80	13.34	0.52	0.83
@81	13.26	0.55	0.83
@82	13.19	0.60	0.83
@83	13.41	0.49	0.83
@84	13.88	**0.17**	**0.85**
@85	13.22	0.59	0.83
@86	13.32	0.44	0.83

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@87	13.31	0.45	0.83
@88	13.19	0.55	0.83
@89	13.31	0.51	0.83
@90	13.34	**0.27**	0.84

Tables E-12 to E-16 show that there were 11 items with conspicuous values in the productive English 1 VST. They were inspected. Eight items did not exhibit any obvious issues. Three items (32, 49, 84) were either ambiguous or incorrect, and they will be revised for version 2 of the productive English 1 VST. An additional two items (61, 63) were identified as ambiguous or poorly phrased. They will also be revised. Five items out of 90 distributed across four bands, however, do not threaten the validity or reliability of the present version of the productive English 1 VST. The data analyzed in this section, therefore, provide strong evidence that the present version (Version 1.1) of the productive English 1 VST exhibits good psychometric properties, meeting all professional standards of reliability and validity required in high-stakes testing.

French: Receptive and Productive Vocabulary Size Tests 1

The French 1 Vocabulary Size Test (VST) was modeled after the English Vocabulary Levels Test pioneered by Paul Nation (Nation, 1990). It uses the word frequency list of the Routledge Frequency Dictionary of French (Lonsdale & Le Bras, 2009). The VST measures how many of its words are known. In this chapter, evidence of validity and reliability of both, the receptive and productive French 1 VST is presented. When test takers take the VST, their results are stored anonymously, without collecting any personal or technical information, to improve the VST. In the present report, data collected between April 2019 and March 2021 were analyzed to examine the overall validity and reliability of the receptive and productive French 1 VST and to identify poorly performing items to be revised.

The Receptive French 1 Vocabulary Size Test

The receptive French 1 VST consists of five bands: the most frequent 1,000; 1,001 to 2,000; 2,001 to 3,000; 3,001 to 4,000; and 4,001 to 5,000 words. It includes ten clusters of six words each for each of the five bands. Each band is thus represented by 60 words. These words involve 30 nouns, 18 verbs, and 12 adjectives and are chosen at random from the 1,000 words of a band. Each cluster focuses on one part of speech, e. g., noun.

Each cluster contains six words and three synonyms, paraphrases, or gapped sentences (targets). Three of the six words are keys, i. e., they correspond to the three targets, while three words are additional distractors. For each target, the same six words are presented as multiple-choice options, one of which needs to be selected for each target. Each band, accordingly, consists of 30 items (targets). The maximum score per band is 30, i. e., 3 points per cluster. The maximum composite score for all five bands is 150, i. e., five times 30.

Correct responses were coded as 1 and incorrect responses as 0. Items that were not attempted were left blank. The maximum time allowed for the five-band test is 30 minutes. Tests that were completed in less than five minutes were removed to reduce the number of test takers who were responding only to a few items. Table F-1 shows the descriptive statistics of the receptive French 1 VST.

Table F-1: Descriptive Statistics of the Receptive French 1 Vocabulary Size Test

N	Mean	SE of Mean	Median	Std. Dev.	Minimum	Maximum
621	83.96	1.47	84	36.74	0	150

Table F-1 shows that there were 621 test takers. Total scores ranged from 0 to 150, covering the complete breadth of scores. The mean and the median were close to the middle of the score range (75) with a standard deviation of 37, indicating that there was a broad cross section of ability levels.

To examine the overall reliability of the receptive French 1 VST, Cronbach's alpha between the five bands of the receptive test was calculated. Cronbach's alpha is a measure of consistency, i. e., how consistent the results of all bands are to each other. It is commonly used as a measure of interrater reliability. Because it is a measure of internal consistency, it may be considered a measure of (internal) validity, i. e., it assesses how well different item sets measure the construct. If alpha is high, it may be assumed that all items measure the same construct, in this case receptive vocabulary size. Cronbach's alpha above 0.7 is considered acceptable, above 0.8, it is considered to be good, and above 0.9 very good. Table F-2 shows the number of tests administered, Cronbach's alpha, and the number of items, in this case bands.

Table F-2: Cronbach's Alpha as a Measure of the Validity and Reliability of the Receptive French 1 Vocabulary Size Test

N of Tests	Cronbach's Alpha	N of Items
515	0.93	5

Table F-2 shows that the reliability and internal validity of the receptive French 1 VST is above 0.9 (very good), which supports the claim that it is highly valid and reliable. Note that the number of tests is different from the number of tests in Table F-1, because to calculate alpha across all bands, all bands need to have values. Some test takers only attempted one or more bands but not all five. For these test takers, Cronbach's alpha of all five bands could not be calculated.

To examine the internal consistency of each band, Cronbach's alpha was calculated for each band of 1,000 words. Each band consists of 30 items. Table F-3 shows the number of test takers, Cronbach's alpha, and the number of items for each band of the receptive French 1 VST.

Table F-3: Cronbach's Alpha for Each Band of the Receptive French 1 Vocabulary Size Test

Band	N of Tests	Alpha	N of Items
1	478	0.91	30
2	422	0.91	30
3	388	0.92	30
4	360	0.92	30
5	357	0.92	30
Total	621		

Table F-3 shows that the internal consistency of each band was above 0.9, i. e., very good, for all bands. To examine the goodness-of-fit of each individual item, the following statistics on the relationship between each individual item and all items of a band were calculated: the scale mean if the item was deleted; the corrected item-total correlation; and Cronbach's alpha if the item was deleted. Items below 0.3 in the column *Corrected Item-Total Correlation* do not correlate well with the overall score and, therefore, provide cause for concern (Field, 2018: 605). Items above the overall Cronbach's alpha of

each band in the column ***Cronbach's Alpha if Item Deleted*** are also problematic, because if their removal raises alpha, then they are less reliable than the average item (Field, 2018: 605). Tables F-4 to F-8 show the summary item statistics for each band of the receptive French 1 VST. Cells with misfitting values are set in bold and red.

Table F-4: Summary Item Statistics for the Receptive French 1 VST Band 1

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@1a	23.75	0.60	0.91
@1b	23.78	0.67	0.90
@1c	23.73	0.59	0.91
@2a	23.68	0.36	0.91
@2b	23.84	0.38	0.91
@2c	23.73	0.51	0.91
@3a	23.70	0.39	0.91
@3b	23.95	0.48	0.91
@3c	23.88	0.56	0.91
@4a	23.76	0.63	0.91
@4b	23.68	0.42	0.91
@4c	23.86	0.39	0.91

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@5a	23.75	0.49	0.91
@5b	23.93	0.48	0.91
@5c	23.88	0.59	0.91
@6a	23.78	0.43	0.91
@6b	23.92	0.52	0.91
@6c	23.88	0.52	0.91
@7a	23.67	0.42	0.91
@7b	24.05	0.50	0.91
@7c	23.93	0.35	0.91
@8a	23.81	0.44	0.91
@8b	23.93	0.47	0.91
@8c	23.75	0.56	0.91
@9a	23.82	0.53	0.91
@9b	23.70	0.54	0.91
@9c	23.69	0.52	0.91
@10a	23.85	0.41	0.91
@10b	23.73	0.49	0.91
@10c	23.87	0.55	0.91

Table F-5: Summary Item Statistics for the Receptive French 1 VST Band 2

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@11a	20.91	0.35	0.91
@11b	20.96	0.52	0.91
@11c	21.14	0.45	0.91
@12a	21.16	0.51	0.91
@12b	21.23	0.40	0.91
@12c	21.08	0.55	0.91
@13a	20.98	0.62	0.91
@13b	21.10	0.56	0.91
@13c	21.04	0.53	0.91
@14a	21.18	0.52	0.91
@14b	21.05	0.60	0.91
@14c	21.28	0.52	0.91
@15a	21.06	0.60	0.91
@15b	21.07	0.56	0.91
@15c	21.04	0.59	0.91
@16a	21.08	0.45	0.91
@16b	21.25	0.63	0.91

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@16c	21.05	0.62	0.91
@17a	21.66	**0.08**	**0.92**
@17b	21.28	0.45	0.91
@17c	21.27	0.51	0.91
@18a	21.17	0.55	0.91
@18b	21.52	0.43	0.91
@18c	21.56	0.36	0.91
@19a	21.03	0.53	0.91
@19b	20.96	0.49	0.91
@19c	20.99	0.55	0.91
@20a	20.95	0.34	0.91
@20b	21.10	0.56	0.91
@20c	20.88	**0.24**	0.91

Table F-6: Summary Item Statistics for the Receptive French 1 VST Band 3

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@21a	16.56	0.54	0.92
@21b	16.64	0.64	0.92

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@21c	16.89	0.64	0.92
@22a	16.84	0.62	0.92
@22b	16.41	0.33	0.92
@22c	16.67	0.55	0.92
@23a	16.61	0.48	0.92
@23b	16.86	0.50	0.92
@23c	16.88	0.69	0.92
@24a	16.55	0.52	0.92
@24b	16.41	0.37	0.92
@24c	16.52	0.52	0.92
@25a	16.65	0.59	0.92
@25b	16.82	0.50	0.92
@25c	16.61	0.38	0.92
@26a	17.08	0.32	0.92
@26b	16.94	0.66	0.92
@26c	16.76	0.69	0.92
@27a	16.81	0.59	0.92
@27b	16.85	0.54	0.92

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@27c	16.89	0.57	0.92
@28a	16.94	0.30	0.92
@28b	16.98	**0.26**	0.92
@28c	16.69	0.55	0.92
@29a	16.78	0.63	0.92
@29b	16.76	0.69	0.92
@29c	16.64	0.37	0.92
@30a	16.73	0.39	0.92
@30b	16.37	**0.24**	0.92
@30c	16.67	0.62	0.92

Table F-7: Summary Item Statistics for the Receptive French 1 VST Band 4

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@31a	18.64	0.54	0.91
@31b	18.59	0.54	0.91
@31c	18.76	0.56	0.91
@32a	18.96	0.36	0.92
@32b	19.05	0.43	0.92

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@32c	18.93	0.57	0.91
@33a	18.46	0.39	0.92
@33b	18.59	0.59	0.91
@33c	18.71	0.58	0.91
@34a	18.94	0.47	0.92
@34b	18.64	0.55	0.91
@34c	18.98	0.51	0.91
@35a	18.71	0.60	0.91
@35b	18.81	0.64	0.91
@35c	18.60	0.30	0.92
@36a	18.55	0.51	0.91
@36b	18.65	0.51	0.91
@36c	18.59	0.50	0.91
@37a	18.79	0.55	0.91
@37b	18.57	0.58	0.91
@37c	18.54	0.48	0.92
@38a	18.54	0.54	0.91
@38b	18.86	0.52	0.91

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@38c	18.71	0.33	0.92
@39a	19.00	0.30	0.92
@39b	19.13	0.34	0.92
@39c	18.69	0.48	0.91
@40a	18.89	0.56	0.91
@40b	18.77	0.62	0.91
@40c	18.72	0.54	0.91

Table F-8: Summary Item Statistics for the Receptive French 1 VST Band 5

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@41a	16.68	0.44	0.92
@41b	16.16	0.35	0.92
@41c	16.52	0.58	0.91
@42a	16.21	0.41	0.92
@42b	16.87	0.35	0.92
@42c	16.62	0.63	0.91
@43a	16.27	0.53	0.91
@43b	16.25	0.35	0.92

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@43c	16.18	**0.27**	0.92
@44a	16.39	0.45	0.92
@44b	16.56	0.63	0.91
@44c	16.72	0.41	0.92
@45a	16.34	0.48	0.92
@45b	16.57	0.64	0.91
@45c	16.64	0.45	0.92
@46a	16.48	0.63	0.91
@46b	16.45	0.63	0.91
@46c	16.39	0.46	0.92
@47a	16.43	0.58	0.91
@47b	16.45	0.48	0.92
@47c	16.70	0.62	0.91
@48a	16.43	0.52	0.91
@48b	16.78	0.40	0.92
@48c	16.70	0.52	0.91
@49a	16.33	0.41	0.92
@49b	16.84	0.48	0.92

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@49c	16.46	0.64	0.91
@50a	16.69	0.49	0.92
@50b	16.45	0.57	0.91
@50c	16.62	0.41	0.92

Tables F-4 to F-8 show that there were only five items in the receptive French 1 Vocabulary Size Test with conspicuous values. They will be inspected and revised, if needed, for version 2 of the receptive French 1 VST. Five items out of 150, however, do not threaten the validity or reliability of the present version of the receptive French 1 VST. The data analyzed in this section, therefore, provide strong evidence that the present version (French 1.1) of the receptive French 1 VST exhibits excellent psychometric properties, meeting all professional standards of reliability and validity required in high-stakes testing.

The Productive French 1 Vocabulary Size Test

The productive vocabulary size test consists of 18 sentences, one for each targeted word, which is partially missing. Each band is thus represented by 18 words. These words involve 9 nouns, 6 verbs, and 3 adjectives chosen at random from the 1,000 words of a band. The maximum score per band is 18. The

maximum composite score for all five bands is 90, i. e., five times 18.

The targeted words appear towards the end of the sentence to establish their meaning. The first few letters are given to disambiguate the word from other possible words. As many letters are provided as needed to disambiguate any given word, up to 50 % of the letters of the word, i. e., if a word consists of an odd number of letters, a maximum of half of the letters minus one are provided. All words of a particular sentence are part of the same band or a more frequent band. In Bands 1 and 2, the partially missing words are uninflected: verbs, for example, appear in their infinitive form. In Bands 3, 4, and 5, partially missing words may be inflected, i. e., grammatical knowledge may be required. Words are scored correct only if they are 100 % correct, including orthographic and grammatical correctness.

Correct responses were coded as 1 and incorrect responses as 0. Items that were not attempted were left blank. Table F–9 shows the descriptive statistics of the productive French 1 VST. Tests that were completed in less than 5 minutes were removed.

Table F–9: Descriptive Statistics of the Productive French 1 Vocabulary Size Test

N	Mean	SE of Mean	Median	Std. Dev.	Minimum	Maximum
190	30.13	1.42	28.00	19.52	0	81

Table F-9 shows that there were 190 test takers. Total scores ranged from 0 to 81, not quite covering the complete breadth of scores. The mean and the median were considerably below the midpoint of the scale (45) with a standard deviation of 20, indicating that most test takers had relatively low vocabulary sizes.

To examine the overall reliability of the productive French 1 VST, Cronbach's alpha between the five bands of the productive test was calculated. Cronbach's alpha is a measure of consistency, i. e., how consistent the results of all bands are to each other. Because it is a measure of internal consistency, it may be considered a measure of (internal) validity, i. e., it assesses how well different item sets measure the construct. If alpha is high, it may be assumed that all items measure the same construct, in this case productive vocabulary size. Cronbach's alpha above 0.7 is considered acceptable, above 0.8, it is considered to be good, and above 0.9 very good. Table F-10 shows the number of tests administered, Cronbach's alpha, and the number of items, in this case bands.

Table F-10: Cronbach's Alpha as a Measure of the Validity and Reliability of the Productive French 1 Vocabulary Size Test

N of Tests	Cronbach's Alpha	N of Items
117	0.91	5

Table F-10 shows that the reliability and internal validity of the productive French 1 VST is above 0.9 (very good), which supports the claim that it is highly valid and reliable. Note that

the number of tests is different from the number of tests in Table F-9, because to calculate alpha across all bands, all bands need to have values. Some test takers only attempted one or more bands but not all five. For these test takers, Cronbach's alpha of all five bands could not be calculated.

To examine the internal consistency of each band, Cronbach's alpha was calculated for each band of 1,000 words. Each band consists of 18 items. Table F-11 shows the number of test takers, Cronbach's alpha, and the number of items for each band of the productive French 1 VST.

Table F-11: Cronbach's Alpha for Each Band of the Productive French 1 Vocabulary Size Test

Band	N of Tests	Alpha	N of Items
1	69	0.65	18
2	40	0.68	18
3	18	0.90	18
4	25	0.87	18
5	22	0.81	18
Total	190		

Table F-11 shows that the internal consistency of Bands 1 and 2 were acceptable (close to 0.7), Band 3 was very good (above 0.9), and Bands 4 and 5 were good (above 0.8). The relatively low alpha at the lower bands may have to do with the relatively low vocabulary sizes of the majority of test takers.

To examine the goodness-of-fit of each individual item, the following statistics on the relationship between each individual item and all items of a band were calculated: the scale mean if the item was deleted; the corrected item-total correlation; and Cronbach's alpha if the item was deleted. Items below 0.3 in the column ***Corrected Item-Total Correlation*** do not correlate well with the overall score and may, therefore, provide cause for concern (Field, 2018: 605). Items above the overall Cronbach's alpha of each band in the column ***Cronbach's Alpha if Item Deleted*** are also problematic, because if their removal raises alpha, then they are less reliable than the average item (Field, 2018: 605). Tables F-12 to F-16 show the summary item statistics for each band of the productive French 1 VST. Cells with misfitting values are set in bold and red.

Table F-12: Summary Item Statistics for the Productive French 1 VST Band 1

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@1	13.32	**0.21**	0.64
@2	13.30	**0.11**	0.65
@3	13.42	0.33	0.63
@4	13.36	**0.05**	**0.66**
@5	13.84	0.45	0.61
@6	13.45	**0.23**	0.64

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@7	13.57	**0.13**	**0.66**
@8	13.26	**0.00**	0.65
@9	13.51	0.40	0.62
@10	13.46	0.44	0.61
@11	13.51	**0.25**	0.64
@12	13.52	0.30	0.63
@13	13.30	**0.17**	0.65
@14	13.70	0.31	0.63
@15	13.39	**0.24**	0.64
@16	13.35	**0.17**	0.65
@17	13.33	**0.24**	0.64
@18	13.84	**0.23**	0.64

Table F-13: Summary Item Statistics for the Productive French 1 VST Band 2

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@19	12.98	**0.16**	0.68
@20	13.08	0.46	0.64
@21	12.77	**0.01**	**0.69**

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@22	12.98	0.42	0.65
@23	12.70	**0.00**	0.68
@24	12.73	**0.16**	0.68
@25	12.70	**0.00**	0.68
@26	13.10	**0.22**	0.68
@27	13.08	0.40	0.65
@28	12.73	**0.21**	0.68
@29	13.08	**0.17**	0.68
@30	13.08	0.40	0.65
@31	12.85	**0.23**	0.67
@32	12.95	0.36	0.66
@33	13.00	**0.27**	0.67
@34	13.18	**0.29**	0.67
@35	12.93	**0.25**	0.67
@36	13.02	0.46	0.64

Table F-14: Summary Item Statistics for the Productive French 1 VST Band 3

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@37	10.89	0.63	0.89
@38	11.28	0.74	0.88
@39	10.83	0.58	0.89
@40	11.28	0.66	0.89
@41	10.94	0.60	0.89
@42	11.44	0.46	0.89
@43	11.33	0.69	0.89
@44	11.17	0.75	0.88
@45	11.22	0.64	0.89
@46	11.17	0.39	0.90
@47	10.83	0.58	0.89
@48	10.89	0.55	0.89
@49	11.11	0.50	0.89
@50	11.56	0.51	0.89
@51	11.22	0.78	0.88
@52	11.06	**0.01**	**0.91**
@53	11.06	0.52	0.89

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@54	10.94	0.36	0.90

Table F-15: Summary Item Statistics for the Productive French 1 VST Band 4

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@55	10.40	0.71	0.85
@56	10.56	**0.26**	0.87
@57	10.40	0.30	0.87
@58	10.60	0.44	0.87
@59	10.24	0.44	0.87
@60	10.52	0.55	0.86
@61	10.72	0.57	0.86
@62	10.44	0.80	0.85
@63	10.20	0.37	0.87
@64	10.60	0.69	0.86
@65	10.68	0.59	0.86
@66	10.68	0.59	0.86
@67	10.28	0.40	0.87
@68	10.08	0.47	0.87

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@69	10.40	0.42	0.87
@70	10.16	0.49	0.86
@71	10.44	0.31	0.87
@72	10.28	0.43	0.87

Table F-16: Summary Item Statistics for the Productive French 1 VST Band 5

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@73	7.45	0.70	0.78
@74	7.59	**0.19**	0.81
@75	7.36	0.32	0.81
@76	7.45	0.47	0.80
@77	7.95	**0.00**	0.81
@78	7.23	0.73	0.78
@79	7.00	**0.18**	0.81
@80	7.82	0.35	0.80
@81	7.36	**0.19**	0.81
@82	7.91	**0.06**	0.81
@83	7.41	0.59	0.79

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@84	7.36	0.40	0.80
@85	7.82	**0.20**	0.81
@86	7.55	**0.26**	0.81
@87	7.36	0.52	0.79
@88	7.27	0.59	0.79
@89	7.50	0.56	0.79
@90	7.82	0.43	0.80

Tables F-12 to F-16 show that there were 24 items with conspicuous values in the productive French 1 VST in Bands 1 and 2 and 9 in Bands 3-5. They will be inspected and revised, where needed, for version 2 of the productive French 1 VST. The probability that many, if not most, of these items will not exhibit any obvious issues is relatively high. Meanwhile, the data analyzed in this section provide sufficient evidence that the present version (Version 1.1) of the productive French 1 VST exhibits acceptable psychometric properties, meeting many professional standards of reliability and validity required in high-stakes testing.

German: Receptive and Productive Vocabulary Size Tests 1

The German 1 Vocabulary Size Test (VST) was modeled after the English Vocabulary Levels Test pioneered by Paul Nation (Nation, 1990). It uses the word frequency list of the Routledge Frequency Dictionary of German (Jones & Tschirner, 2006). The VST measures how many of its words are known. In this chapter, evidence of validity and reliability of both, the receptive and productive German 1 VST is presented. When test takers take the VST, their results are stored anonymously, without collecting any personal or technical information, to improve the VST. In the present report, data collected between April 2019 and March 2021 were analyzed to examine the overall validity and reliability of the receptive and productive German 1 VST and to identify poorly performing items to be revised.

The Receptive German 1 Vocabulary Size Test

The receptive German 1 VST consists of five bands: the most frequent 1,000; 1,001 to 2,000; 2,001 to 3,000; 3,001 to 4,000; and 4,001 to 5,000 words. It includes ten clusters of six words each for each of the five bands. Each band is thus represented by 60 words. These words involve 30 nouns, 18 verbs, and 12 adjectives and are chosen at random from the 1,000 words of a band. Each cluster focuses on one part of speech, e. g., noun.

Each cluster contains six words and three synonyms, paraphrases, or gapped sentences (targets). Three of the six words are keys, i. e., they correspond to the three targets, while three words are additional distractors. For each target, the same six words are presented as multiple-choice options, one of which needs to be selected for each target. Each band, accordingly, consists of 30 items (targets). The maximum score per band is 30, i. e., 3 points per cluster. The maximum composite score for all five bands is 150, i. e., five times 30.

Correct responses were coded as 1 and incorrect responses as 0. Items that were not attempted were left blank. The maximum time allowed for the five-band test is 30 minutes. Tests that were completed in less than five minutes were removed to reduce the number of test takers who were responding only to a few items. Table G-1 shows the descriptive statistics of the receptive German 1 VST.

Table G-1: Descriptive Statistics of the Receptive German 1 Vocabulary Size Test

N	Mean	SE of Mean	Median	Std. Dev.	Minimum	Maximum
2214	97.25	0.98	106	45.86	0	150

Table G-1 shows that there were 2214 test takers. Total scores ranged from 0 to 150, covering the complete breadth of scores. The mean and the median were close to or within the upper third of the score range, indicating that there was a large number of test takers with high receptive vocabulary sizes.

To examine the overall reliability of the receptive German 1 VST, Cronbach's alpha between the five bands of the receptive test was calculated. Cronbach's alpha is a measure of consistency, i. e., how consistent the results of all bands are to each other. It is commonly used as a measure of interrater reliability. Because it is a measure of internal consistency, it may be considered a measure of (internal) validity, i. e., it assesses how well different item sets measure the construct. If alpha is high, it may be assumed that all items measure the same construct, in this case receptive vocabulary size. Cronbach's alpha above 0.7 is considered acceptable, above 0.8, it is considered to be good, and above 0.9 very good. Table G-2 shows the number of tests administered, Cronbach's alpha, and the number of items, in this case bands.

Table G-2: Cronbach's Alpha as a Measure of the Validity and Reliability of the Receptive German 1 Vocabulary Size Test

N of Tests	Cronbach's Alpha	N of Items
1577	0.94	5

Table G-2 shows that the reliability and internal validity of the receptive German 1 VST was above 0.9 (very good), which supports the claim that it is highly valid and reliable. Note that the number of tests is different from the number of tests in Table G-1, because to calculate alpha across all bands, all bands need to have values. Some test takers only attempted one or more bands but not all five. For these test takers, Cronbach's alpha of all five bands could not be calculated.

To examine the internal consistency of each band, Cronbach's alpha was calculated for each band of 1,000 words. Each band consists of 30 items. Table G-3 shows the number of test takers, Cronbach's alpha, and the number of items for each band of the receptive German 1 VST.

Table G-3: Cronbach's Alpha for Each Band of the Receptive German 1 Vocabulary Size Test

Band	N of Tests	Alpha	N of Items
1	1704	0.88	30
2	1594	0.94	30
3	1415	0.94	30
4	1251	0.94	30
5	1123	0.94	30
Total	2238		

Table G-3 shows that the internal consistency of each band was close to or above 0.9, i. e., close to or within the very good range. To examine the goodness-of-fit of each individual item, the following statistics on the relationship between each individual item and all items of a band were calculated: the scale mean if the item was deleted; the corrected item-total correlation; and Cronbach's alpha if the item was deleted. Items below 0.3 in the column *Corrected Item-Total Correlation* do not correlate well with the overall score and, therefore, provide cause for concern (Field, 2018: 605). Items above the

overall Cronbach's alpha of each band in the column ***Cronbach's Alpha if Item Deleted*** are also problematic, because if their removal raises alpha, then they are less reliable than the average item (Field, 2018: 605). Tables G-4 to G-8 show the summary item statistics for each band of the receptive German 1 VST. Cells with misfitting values are set in bold and red.

Table G-4: Summary Item Statistics for the Receptive German 1 VST Band 1

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@1a	26.66	0.49	0.88
@1b	26.61	**0.22**	0.88
@1c	26.58	0.47	0.88
@2a	26.56	**0.29**	0.88
@2b	26.60	0.46	0.88
@2c	26.68	0.54	0.88
@3a	26.64	0.46	0.88
@3b	26.57	0.46	0.88
@3c	26.60	0.46	0.88
@4a	26.62	0.45	0.88
@4b	26.58	0.43	0.88
@4c	26.60	0.33	0.88

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@5a	26.60	0.43	0.88
@5b	26.66	0.44	0.88
@5c	26.63	0.45	0.88
@6a	26.86	0.33	**0.89**
@6b	26.61	0.44	0.88
@6c	26.65	0.58	0.88
@7a	26.61	0.41	0.88
@7b	26.61	0.50	0.88
@7c	26.73	0.50	0.88
@8a	26.58	0.43	0.88
@8b	26.61	0.53	0.88
@8c	26.65	0.50	0.88
@9a	26.58	0.36	0.88
@9b	26.74	0.37	0.88
@9c	26.61	0.41	0.88
@10a	26.60	0.52	0.88
@10b	26.77	0.50	0.88
@10c	26.59	0.50	0.88

Table G-5: Summary Item Statistics for the Receptive German 1 VST Band 2

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@11a	25.07	0.65	0.94
@11b	25.04	0.49	0.94
@11c	24.97	0.38	0.94
@12a	25.12	0.61	0.94
@12b	25.07	0.65	0.94
@12c	25.02	0.57	0.94
@13a	25.02	0.53	0.94
@13b	25.13	0.65	0.94
@13c	25.04	0.63	0.94
@14a	25.03	0.50	0.94
@14b	25.03	0.51	0.94
@14c	25.07	0.63	0.94
@15a	25.10	0.52	0.94
@15b	24.98	0.44	0.94
@15c	25.06	0.64	0.94
@16a	25.10	0.55	0.94
@16b	25.20	0.64	0.94

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@16c	25.10	0.55	0.94
@17a	25.00	0.46	0.94
@17b	25.09	0.57	0.94
@17c	25.09	0.52	0.94
@18a	25.02	0.56	0.94
@18b	25.06	0.56	0.94
@18c	25.11	0.55	0.94
@19a	25.09	0.53	0.94
@19b	25.14	0.63	0.94
@19c	25.20	0.59	0.94
@20a	25.07	0.66	0.94
@20b	25.06	0.58	0.94
@20c	25.07	0.64	0.94

Table G-6: Summary Item Statistics for the Receptive German 1 VST Band 3

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@21a	23.85	0.62	0.93
@21b	23.74	0.41	0.93

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@21c	23.84	0.58	0.93
@22a	23.77	0.43	0.93
@22b	23.78	0.57	0.93
@22c	23.78	0.58	0.93
@23a	23.78	0.57	0.93
@23b	23.78	0.58	0.93
@23c	23.84	0.57	0.93
@24a	23.79	0.50	0.93
@24b	23.96	0.66	0.93
@24c	23.91	0.68	0.93
@25a	23.93	0.55	0.93
@25b	23.79	0.53	0.93
@25c	23.92	0.60	0.93
@26a	23.83	0.61	0.93
@26b	23.93	0.67	0.93
@26c	23.95	0.67	0.93
@27a	23.79	0.55	0.93
@27b	23.96	0.62	0.93

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@27c	23.91	0.68	0.93
@28a	24.20	**0.21**	**0.94**
@28b	23.80	0.50	0.93
@28c	23.90	0.46	0.93
@29a	23.94	0.54	0.93
@29b	23.92	0.45	0.93
@29c	23.78	0.48	0.93
@30a	23.77	0.57	0.93
@30b	23.78	0.50	0.93
@30c	23.91	0.59	0.93

Table G-7: Summary Item Statistics for the Receptive German 1 VST Band 4

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@31a	23.65	**0.29**	0.94
@31b	23.76	0.55	0.94
@31c	23.80	0.59	0.94
@32a	23.75	0.55	0.94
@32b	23.81	0.62	0.94

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@32c	23.75	0.57	0.94
@33a	23.79	0.62	0.94
@33b	23.92	0.67	0.94
@33c	23.73	0.57	0.94
@34a	23.71	0.54	0.94
@34b	23.75	0.53	0.94
@34c	23.67	0.45	0.94
@35a	23.74	0.53	0.94
@35b	23.81	0.58	0.94
@35c	23.91	0.68	0.94
@36a	23.77	0.57	0.94
@36b	23.98	0.64	0.94
@36c	23.79	0.53	0.94
@37a	23.99	0.67	0.94
@37b	23.95	0.68	0.94
@37c	23.84	0.64	0.94
@38a	23.86	0.66	0.94
@38b	23.96	0.70	0.94

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@38c	23.70	0.52	0.94
@39a	23.73	0.47	0.94
@39b	23.79	0.51	0.94
@39c	23.73	0.57	0.94
@40a	23.90	0.67	0.94
@40b	24.02	0.47	0.94
@40c	23.73	0.58	0.94

Table G-8: Summary Item Statistics for the Receptive German 1 VST Band 5

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@41a	24.23	0.59	0.94
@41b	24.13	0.44	0.94
@41c	24.51	0.33	0.94
@42a	24.23	0.58	0.94
@42b	24.27	0.64	0.94
@42c	24.26	0.57	0.94
@43a	24.27	0.62	0.94
@43b	24.15	0.50	0.94

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@43c	24.23	0.66	0.94
@44a	24.20	0.58	0.94
@44b	24.24	0.68	0.94
@44c	24.16	0.58	0.94
@45a	24.14	0.51	0.94
@45b	24.18	0.63	0.94
@45c	24.15	0.60	0.94
@46a	24.38	0.72	0.94
@46b	24.23	0.66	0.94
@46c	24.29	0.65	0.94
@47a	24.26	0.72	0.94
@47b	24.20	0.68	0.94
@47c	24.18	0.57	0.94
@48a	24.36	0.49	0.94
@48b	24.34	0.59	0.94
@48c	24.22	0.65	0.94
@49a	24.19	0.62	0.94
@49b	24.19	0.67	0.94

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@49c	24.23	0.58	0.94
@50a	24.21	0.52	0.94
@50b	24.41	0.60	0.94
@50c	24.46	0.44	0.94

Tables G-4 to G-8 show that there were only five items in the receptive German 1 Vocabulary Size Test with conspicuous values. They were inspected. Three of these items (1b, 2a, 31a) had slightly different difficulty values than the average item of their respective band but did not exhibit any additional issues. These items will not be revised. Two items (6a, 28a) were found to be incorrect and will be revised for version 2 of the receptive German 1 VST. Two items out of 150, however, do not threaten the validity or reliability of the present version of the receptive German 1 VST.

The data analyzed in this section, therefore, provide strong evidence that the present version (German 1.1) of the receptive German 1 VST exhibits excellent psychometric properties, meeting all professional standards of reliability and validity required in high-stakes testing.

The Productive German 1 Vocabulary Size Test

The productive vocabulary size test consists of 18 sentences, one for each targeted word, which is partially missing. Each

band is thus represented by 18 words. These words involve 9 nouns, 6 verbs, and 3 adjectives chosen at random from the 1,000 words of a band. The maximum score per band is 18. The maximum composite score for all five bands is 90, i. e., five times 18.

The targeted words appear towards the end of the sentence to establish their meaning. The first few letters are given to disambiguate the word from other possible words. As many letters are provided as needed to disambiguate any given word, up to 50 % of the letters of the word, i. e., if a word consists of an odd number of letters, a maximum of half of the letters minus one are provided. All words of a particular sentence are part of the same band or a more frequent band. In Bands 1 and 2, the partially missing words are uninflected: verbs, for example, appear in their infinitive form. In Bands 3, 4, and 5, partially missing words may be inflected, i. e., grammatical knowledge may be required. Words are scored correct only if they are 100 % correct, including orthographic and grammatical correctness.

Correct responses were coded as 1 and incorrect responses as 0. Items that were not attempted were left blank. Table G–9 shows the descriptive statistics of the productive German 1 VST. Tests that were completed in less than 5 minutes were removed.

Table G-9: Descriptive Statistics of the Productive German 1 Vocabulary Size Test

N	Mean	SE of Mean	Median	Std. Dev.	Minimum	Maximum
1891	54.09	0.63	61	27.51	0	88

Table G-9 shows that there were 1,891 test takers. Total scores ranged from 0 to 88, all but covering the complete breadth of scores. The mean and the median were above the midpoint of the scale (45) with a standard deviation of 27.5, indicating that there was a broad range of test takers.

To examine the overall reliability of the productive German 1 VST, Cronbach's alpha between the five bands of the productive test was calculated. Cronbach's alpha is a measure of consistency, i. e., how consistent the results of all bands are to each other. Because it is a measure of internal consistency, it may be considered a measure of (internal) validity, i. e., it assesses how well different item sets measure the construct. If alpha is high, it may be assumed that all items measure the same construct, in this case productive vocabulary size. Cronbach's alpha above 0.7 is considered acceptable, above 0.8, it is considered to be good, and above 0.9 very good. Table G-10 shows the number of tests administered, Cronbach's alpha, and the number of items, in this case bands.

Table G-10: Cronbach's Alpha as a Measure of the Validity and Reliability of the Productive German 1 Vocabulary Size Test

N of Tests	Cronbach's Alpha	N of Items
1551	0.95	5

Table G-10 shows that the reliability and internal validity of the productive German 1 VST is considerably above 0.9 (very good), which supports the claim that it is highly valid and reliable. Note that the number of tests is different from the number of tests in Table G-9, because to calculate alpha across all bands, all bands need to have values. Some test takers only attempted one or more bands but not all five. For these test takers, Cronbach's alpha of all five bands could not be calculated.

To examine the internal consistency of each band, Cronbach's alpha was calculated for each band of 1,000 words. Each band consists of 18 items. Table G-11 shows the number of test takers, Cronbach's alpha, and the number of items for each band of the productive German 1 VST.

Table G-11: Cronbach's Alpha for Each Band of the Productive German 1 Vocabulary Size Test

Band	N of Tests	Alpha	N of Items
1	1131	0.84	18
2	891	0.79	18
3	765	0.77	18

Band	N of Tests	Alpha	N of Items
4	733	0.78	18
5	806	0.87	18
Total	2238		

Table G-11 shows that the internal consistency of each band was close to or above 0.8, i. e., close to or within the good range.

To examine the goodness-of-fit of each individual item, the following statistics on the relationship between each individual item and all items of a band were calculated: the scale mean if the item was deleted; the corrected item-total correlation; and Cronbach's alpha if the item was deleted. Items below 0.3 in the column ***Corrected Item-Total Correlation*** do not correlate well with the overall score and may, therefore, provide cause for concern (Field, 2018: 605). Items above the overall Cronbach's alpha of each band in the column ***Cronbach's Alpha if Item Deleted*** are also problematic, because if their removal raises alpha, then they are less reliable than the average item (Field, 2018: 605). Tables G-12 to G-16 show the summary item statistics for each band of the productive German 1 VST. Cells with misfitting values are set in bold and red.

Table G-12: Summary Item Statistics for the Productive German 1 VST Band 1

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@1	15.57	**0.26**	0.84
@2	15.55	0.40	0.83
@3	15.52	0.45	0.83
@4	15.54	0.32	0.84
@5	15.65	0.58	0.82
@6	15.52	0.44	0.83
@7	15.52	**0.25**	0.84
@8	15.58	0.38	0.84
@9	15.57	0.51	0.83
@10	15.59	0.48	0.83
@11	15.54	0.44	0.83
@12	15.58	0.44	0.83
@13	15.62	0.54	0.83
@14	15.62	0.41	0.83
@15	15.67	0.55	0.83
@16	15.55	0.41	0.83
@17	15.60	0.54	0.83

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@18	15.63	0.49	0.83

Table G-13: Summary Item Statistics for the Productive German 1 VST Band 2

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@19	15.21	0.30	0.79
@20	15.13	**0.16**	0.79
@21	15.13	0.39	0.78
@22	15.19	0.38	0.78
@23	15.17	0.39	0.78
@24	15.41	0.44	0.78
@25	15.33	**0.27**	0.79
@26	15.22	0.54	0.77
@27	15.19	0.48	0.78
@28	15.21	0.55	0.77
@29	15.13	**0.29**	0.79
@30	15.19	0.42	0.78
@31	15.20	0.44	0.78
@32	15.17	0.51	0.78

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@33	15.18	0.30	0.79
@34	15.12	**0.13**	0.79
@35	15.38	0.41	0.78
@36	15.15	0.40	0.78

Table G-14: Summary Item Statistics for the Productive German 1 VST Band 3

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@37	14.46	**0.24**	0.77
@38	14.37	0.34	0.77
@39	14.40	0.50	0.76
@40	14.42	0.47	0.76
@41	14.49	**0.27**	0.77
@42	14.49	0.48	0.75
@43	14.41	0.52	0.75
@44	14.45	0.48	0.75
@45	14.44	**0.01**	**0.79**
@46	14.38	0.34	0.77
@47	14.43	0.39	0.76

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@48	14.47	0.37	0.76
@49	14.53	**0.29**	0.77
@50	14.52	0.51	0.75
@51	14.56	0.48	0.75
@52	15.09	**0.11**	0.79
@53	14.50	0.49	0.75
@54	14.50	0.32	0.77

Table G-15: Summary Item Statistics for the Productive German 1 VST Band 4

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@55	14.69	0.36	0.77
@56	14.78	0.50	0.76
@57	14.72	**0.24**	0.77
@58	14.91	**0.28**	0.78
@59	14.84	**0.23**	0.78
@60	14.81	0.31	0.77
@61	14.75	0.42	0.76
@62	14.74	0.43	0.76

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@63	14.73	0.57	0.76
@64	14.80	0.46	0.76
@65	14.80	**0.19**	0.78
@66	14.71	0.51	0.76
@67	14.73	0.47	0.76
@68	14.72	0.47	0.76
@69	14.79	0.42	0.76
@70	14.86	0.44	0.76
@71	14.97	0.33	0.77
@72	14.84	**0.15**	0.78

Table G-16: Summary Item Statistics for the Productive German 1 VST Band 5

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@73	14.94	0.62	0.85
@74	14.89	0.35	0.86
@75	14.93	0.59	0.85
@76	14.97	0.46	0.86
@77	14.92	0.60	0.85

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@78	14.94	0.51	0.86
@79	14.93	0.53	0.86
@80	15.54	**0.05**	**0.89**
@81	14.92	0.60	0.85
@82	14.91	0.56	0.86
@83	14.91	0.59	0.85
@84	14.93	0.66	0.85
@85	14.90	0.59	0.86
@86	14.85	0.43	0.86
@87	14.96	0.64	0.85
@88	14.86	0.37	0.86
@89	14.93	0.43	0.86
@90	14.89	0.35	0.86

Tables G-12 to G-16 show that there were 17 items with conspicuous values in the productive German 1 VST. They were inspected. Most of these items were either somewhat easier than the average item of their respective band (7, 20, 29, 34, 41, 45) or somewhat more difficult (25, 49, 52, 57, 58, 59, 65) but did not exhibit any additional issues. Four items (7, 37, 72, 80) were either ambiguous or poorly phrased and will be revised for version 2 of the productive German 1 VST. Four items out of 90,

however, do not threaten the validity or reliability of the present version of the productive German 1 VST. The data analyzed in this section, therefore, provide strong evidence that the present version (Version 1.1) of the productive German 1 VST exhibits good psychometric properties, meeting all professional standards of reliability and validity required in high-stakes testing.

Italian: Receptive Vocabulary Size Test 1

The Italian 1 Vocabulary Size Test (VST) was modeled after the English Vocabulary Levels Test pioneered by Paul Nation (Nation, 1990). It is based on De Mauro et al. (1993) and De Mauro et al. (1996) and measures how many of the most frequent 5,000 words of Italian are known. In this chapter, evidence of validity and reliability of the receptive Italian 1 VST is presented. Not enough data were available for the productive Italian 1 VST (less than 30 datasets).

The receptive Italian 1 VST consists of five bands: the most frequent 1,000; 1,001 to 2,000; 2,001 to 3,000; 3,001 to 4,000; and 4,001 to 5,000 words. It includes ten clusters of six words each for each of the five bands. Each band is thus represented by 60 words. These words involve 30 nouns, 18 verbs, and 12 adjectives and are chosen at random from the 1,000 words of a band. Each cluster focuses on one part of speech, e. g., noun.

Each cluster contains six words and three synonyms, paraphrases, or gapped sentences (targets). Three of the six words are keys, i. e., they correspond to the three targets, while three words are additional distractors. For each target, the same six words are presented as multiple-choice options, one of which needs to be selected for each target. Each band, accordingly, consists of 30 items (targets). The maximum score per band is 30, i. e., 3 points per cluster. The maximum composite score for all five bands is 150, i. e., five times 30.

When test takers take the VST, their results are stored anonymously, without collecting any personal or technical information, to improve the VST. In the present report, data collected between April 2019 and March 2021 were analyzed to examine the overall validity and reliability of the receptive Italian 1 VST and to identify poorly performing items to be revised.

Correct responses were coded as 1 and incorrect responses as 0. Items that were not attempted were left blank. The maximum time allowed for the five-band test is 30 minutes. Tests that were completed in less than five minutes were removed to reduce the number of test takers who were responding only to a few items. Table I-1 shows the descriptive statistics of the receptive Italian 1 VST.

Table I-1: Descriptive Statistics of the Receptive Italian 1 Vocabulary Size Test

N	Mean	SE of Mean	Median	Std. Dev.	Minimum	Maximum
76	92.17	4.73	104.50	41.25	8	145

Table I-1 shows that there were 76 test takers. Total scores ranged from 8 to 145, covering most of the scores. The mean and the median were close to or within the upper third of the score range with a standard deviation of over 40, indicating that there was a large number of test takers with high receptive vocabulary sizes.

To examine the overall reliability of the receptive Italian 1 VST, Cronbach's alpha between the five bands of the receptive test was calculated. Cronbach's alpha is a measure of consistency, i. e., how consistent the results of all bands are to each other. It is commonly used as a measure of interrater reliability. Because it is a measure of internal consistency, it may be considered a measure of (internal) validity, i. e., it assesses how well different item sets measure the construct. If alpha is high, it may be assumed that all items measure the same construct, in this case receptive vocabulary size. Cronbach's alpha above 0.7 is considered acceptable, above 0.8, it is considered to be good, and above 0.9 very good. Table I-2 shows the number of tests administered, Cronbach's alpha, and the number of items, in this case bands.

Table I-2: Cronbach's Alpha as a Measure of the Validity and Reliability of the Receptive Italian 1 Vocabulary Size Test

N of Tests	Cronbach's Alpha	N of Items
58	0.92	5

Table I-2 shows that the reliability and internal validity of the receptive Italian 1 VST is above 0.9 (very good), which supports the claim that it is highly valid and reliable. Note that the number of tests is different from the number of tests in Table I-1, because to calculate alpha across all bands, all bands need to have values. Some test takers only attempted one or more bands but not all five. For these test takers, Cronbach's alpha of all five bands could not be calculated.

To examine the internal consistency of each band, Cronbach's alpha was calculated for each band of 1,000 words. Each band consists of 30 items. Table I-3 shows the number of test takers, Cronbach's alpha, and the number of items for each band of the receptive Italian 1 VST.

Table I-3: Cronbach's Alpha for Each Band of the Receptive Italian 1 Vocabulary Size Test

Band	N of Tests	Alpha	N of Items
1	40	0.80	30
2	49	0.75	30
3	39	0.90	30
4	31	0.90	30
5	32	0.91	30
Total	76		

Table I-3 shows that the internal consistency of Band 1 was good (0.8), Band 2 was acceptable (above 0.7), and Bands 3-5 were very good (at or above 0.9). To examine the goodness-of-fit of each individual item, the following statistics on the relationship between each individual item and all items of a band were calculated: the scale mean if the item was deleted; the corrected item-total correlation; and Cronbach's alpha if the item was deleted. Items below 0.3 in the column ***Corrected Item-Total Correlation*** do not correlate well with the overall score and, therefore, provide cause for concern (Field, 2018:

605). Items above the overall Cronbach's alpha of each band in the column ***Cronbach's Alpha if Item Deleted*** are also problematic, because if their removal raises alpha, then they are less reliable than the average item (Field, 2018: 605). Tables I-4 to I-8 show the summary item statistics for each band of the receptive Italian 1 VST. Cells with misfitting values are set in bold and red.

Table I-4: Summary Item Statistics for the Receptive Italian 1 VST Band 1

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@1a	26.28	0.66	0.78
@1b	26.30	**0.06**	**0.81**
@1c	26.25	0.58	0.79
@2a	26.28	0.58	0.79
@2b	26.40	**0.06**	**0.81**
@2c	26.25	**0.27**	0.80
@3a	26.48	0.38	0.79
@3b	26.37	0.40	0.79
@3c	26.33	0.30	0.80
@4a	26.28	0.66	0.78
@4b	26.25	0.58	0.79

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@4c	26.25	0.58	0.79
@5a	26.45	**0.16**	**0.81**
@5b	26.37	0.30	0.80
@5c	26.30	0.49	0.79
@6a	26.28	**0.24**	0.80
@6b	26.25	**-0.04**	**0.81**
@6c	26.48	0.34	0.80
@7a	26.23	**0.00**	0.80
@7b	26.45	0.53	0.78
@7c	26.45	**0.14**	**0.81**
@8a	26.42	0.58	0.78
@8b	26.37	0.42	0.79
@8c	26.23	**0.00**	0.80
@9a	26.28	0.50	0.79
@9b	26.25	**-0.09**	**0.81**
@9c	26.23	**0.00**	0.80
@10a	26.25	0.58	0.79
@10b	26.28	**-0.02**	**0.81**

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@10c	26.28	0.50	0.79

Table I-5: Summary Item Statistics for the Receptive Italian 1 VST Band 2

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@11a	26.29	**0.22**	0.74
@11b	26.24	**0.00**	0.75
@11c	26.27	0.45	0.74
@12a	26.29	0.37	0.74
@12b	26.29	0.33	0.74
@12c	26.29	0.41	0.73
@13a	26.29	**0.10**	0.75
@13b	26.27	0.39	0.74
@13c	26.27	0.39	0.74
@14a	26.41	**0.01**	**0.76**
@14b	26.27	**-0.04**	0.75
@14c	26.27	**0.07**	0.75
@15a	26.24	**0.00**	0.75
@15b	26.27	**-0.04**	0.75

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@15c	26.24	**0.00**	0.75
@16a	26.29	**0.29**	0.74
@16b	26.86	**0.16**	**0.76**
@16c	26.51	0.34	0.74
@17a	26.27	0.45	0.74
@17b	26.47	0.39	0.73
@17c	26.35	0.66	0.71
@18a	26.24	**0.00**	0.75
@18b	26.37	0.42	0.73
@18c	26.33	0.66	0.72
@19a	26.47	**0.21**	0.75
@19b	26.29	0.50	0.73
@19c	26.41	**0.20**	0.75
@20a	26.35	0.49	0.72
@20b	26.39	0.36	0.73
@20c	26.37	**0.15**	0.75

Table I-6: Summary Item Statistics for the Receptive Italian 1 VST Band 3

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@21a	23.67	0.47	0.90
@21b	23.59	0.36	0.90
@21c	23.67	0.41	0.90
@22a	23.74	0.70	0.89
@22b	23.64	0.62	0.89
@22c	23.62	0.36	0.90
@23a	23.62	**0.29**	0.90
@23b	23.77	0.58	0.90
@23c	23.64	**0.07**	0.90
@24a	23.64	0.62	0.89
@24b	24.03	0.34	0.90
@24c	23.56	0.52	0.90
@25a	23.74	0.58	0.90
@25b	23.59	0.70	0.89
@25c	23.62	0.62	0.90
@26a	23.62	0.49	0.90
@26b	23.67	0.56	0.90

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@26c	23.67	0.48	0.90
@27a	23.62	**0.19**	0.90
@27b	23.69	0.65	0.89
@27c	23.69	**0.24**	0.90
@28a	23.56	0.45	0.90
@28b	23.62	0.44	0.90
@28c	24.54	**0.00**	0.90
@29a	23.59	0.70	0.89
@29b	23.74	0.78	0.89
@29c	23.72	0.67	0.89
@30a	23.62	0.69	0.89
@30b	23.62	**0.08**	0.90
@30c	24.54	**0.00**	0.90

Table I-7: Summary Item Statistics for the Receptive Italian 1 VST Band 4

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@31a	22.52	0.59	0.90
@31b	22.58	0.76	0.89

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@31c	22.58	0.65	0.90
@32a	22.52	0.59	0.90
@32b	22.52	0.59	0.90
@32c	22.55	0.64	0.90
@33a	22.71	0.42	0.90
@33b	22.65	0.87	0.89
@33c	22.65	0.71	0.89
@34a	22.48	**0.00**	0.90
@34b	22.55	0.32	0.90
@34c	22.52	0.59	0.90
@35a	22.90	0.58	0.90
@35b	22.74	0.70	0.89
@35c	22.71	**0.27**	0.90
@36a	22.84	0.42	0.90
@36b	23.16	0.34	0.90
@36c	22.65	0.76	0.89
@37a	22.87	0.42	0.90
@37b	23.13	0.39	0.90

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@37c	23.00	**-0.01**	**0.91**
@38a	22.90	0.56	0.90
@38b	22.81	0.31	0.90
@38c	23.13	**0.23**	0.90
@39a	22.55	0.64	0.90
@39b	22.61	0.82	0.89
@39c	22.55	0.39	0.90
@40a	22.61	0.44	0.90
@40b	22.52	0.35	0.90
@40c	22.55	0.64	0.90

Table I-8: Summary Item Statistics for the Receptive Italian 1 VST Band 5

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@41a	21.41	0.46	0.91
@41b	21.81	0.41	0.91
@41c	22.28	**0.00**	**0.92**
@42a	21.34	0.66	0.91
@42b	21.31	0.59	0.91

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@42c	21.34	0.66	0.91
@43a	21.38	0.88	0.91
@43b	21.44	0.52	0.91
@43c	21.34	0.81	0.91
@44a	21.28	**0.00**	**0.92**
@44b	21.53	**0.21**	**0.92**
@44c	21.56	0.61	0.91
@45a	21.53	0.61	0.91
@45b	21.47	0.49	0.91
@45c	21.59	0.38	0.91
@46a	21.44	0.48	0.91
@46b	22.28	**0.00**	**0.92**
@46c	21.44	0.76	0.91
@47a	21.38	0.60	0.91
@47b	21.41	0.65	0.91
@47c	21.50	0.64	0.91
@48a	21.50	0.55	0.91
@48b	21.38	0.79	0.91

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@48c	21.53	0.70	0.91
@49a	21.69	0.41	0.91
@49b	21.31	0.52	0.91
@49c	21.31	**-0.06**	**0.92**
@50a	21.34	0.81	0.91
@50b	21.75	0.44	0.91
@50c	22.28	**0.00**	**0.92**

Tables I-4 to I-8 show that there were 27 items with conspicuous values in the receptive Italian 1 VST in Bands 1 and 2 and 17 items in Bands 3-5, i. e., almost a third of all items. They will be inspected and revised, where needed, for version 2 of the receptive Italian 1 VST. The probability that many, if not most, of these items will not exhibit any obvious issues is relatively high. Meanwhile, the data analyzed in this chapter provide sufficient evidence that the present version (Version 1.1) of the receptive Italian 1 VST exhibits acceptable psychometric properties, meeting many professional standards of reliability and validity required in high-stakes testing.

Russian: Receptive Vocabulary Size Test 1

The Italian 1 Vocabulary Size Test (VST) was modeled after the English Vocabulary Levels Test pioneered by Paul Nation (Nation, 1990). It uses the word frequency list of the Routledge Frequency Dictionary of Russian (Sharoff, Umanskaya, & Wilson, 2013), which contains the most frequent 5,000 words of Russian, and measures how many of them are known. In this chapter, evidence of validity and reliability of the receptive Russian 1 VST is presented. Not enough data were available for the productive Russian 1 VST (less than 30 datasets).

The receptive Russian 1 VST consists of five bands: the most frequent 1,000; 1,001 to 2,000; 2,001 to 3,000; 3,001 to 4,000; and 4,001 to 5,000 words. It includes ten clusters of six words each for each of the five bands. Each band is thus represented by 60 words. These words involve 30 nouns, 18 verbs, and 12 adjectives and are chosen at random from the 1,000 words of a band. Each cluster focuses on one part of speech, e. g., noun.

Each cluster contains six words and three synonyms, paraphrases, or gapped sentences (targets). Three of the six words are keys, i. e., they correspond to the three targets, while three words are additional distractors. For each target, the same six words are presented as multiple-choice options, one of which needs to be selected for each target. Each band, accordingly, consists of 30 items (targets). The maximum score per band is 30, i. e., 3 points per cluster. The maximum composite score for all five bands is 150, i. e., five times 30.

When test takers take the VST, their results are stored anonymously, without collecting any personal or technical information, to improve the VST. In the present report, data collected between April 2019 and March 2021 were analyzed to examine the overall validity and reliability of the receptive Russian 1 VST and to identify poorly performing items to be revised.

Correct responses were coded as 1 and incorrect responses as 0. Items that were not attempted were left blank. The maximum time allowed for the five-band test is 30 minutes. Tests that were completed in less than five minutes were removed to reduce the number of test takers who were responding only to a few items. Table R-1 shows the descriptive statistics of the receptive Russian 1 VST.

Table R-1: Descriptive Statistics of the Receptive Russian 1 Vocabulary Size Test

N	Mean	SE of Mean	Median	Std. Dev.	Minimum	Maximum
198	100.98	3.37	115.50	47.40	0	150

Table R-1 shows that there were 198 test takers. Total scores ranged from 0 to 150, covering the complete breadth of scores. The mean and the median were within the upper third of the score range, indicating that there was a large number of test takers with high receptive vocabulary sizes.

To examine the overall reliability of the receptive Russian 1 VST, Cronbach's alpha between the five bands of the receptive

test was calculated. Cronbach's alpha is a measure of consistency, i. e., how consistent the results of all bands are to each other. It is commonly used as a measure of interrater reliability. Because it is a measure of internal consistency, it may be considered a measure of (internal) validity, i. e., it assesses how well different item sets measure the construct. If alpha is high, it may be assumed that all items measure the same construct, in this case receptive vocabulary size. Cronbach's alpha above 0.7 is considered acceptable, above 0.8, it is considered to be good, and above 0.9 very good. Table R-2 shows the number of tests administered, Cronbach's alpha, and the number of items, in this case bands.

Table R-2: Cronbach's Alpha as a Measure of the Validity and Reliability of the Receptive Russian 1 Vocabulary Size Test

N of Tests	Cronbach's Alpha	N of Items
153	0.95	5

Table R-2 shows that the reliability and internal validity of the receptive Russian 1 VST was above 0.9 (very good), which supports the claim that it is highly valid and reliable. Note that the number of tests is different from the number of tests in Table R-1, because to calculate alpha across all bands, all bands need to have values. Some test takers only attempted one or more bands but not all five. For these test takers, Cronbach's alpha of all five bands could not be calculated.

To examine the internal consistency of each band, Cronbach's alpha was calculated for each band of 1,000 words.

Each band consists of 30 items. Table R-3 shows the number of test takers, Cronbach's alpha, and the number of items for each band of the receptive Russian 1 VST.

Table R-3: Cronbach's Alpha for Each Band of the Receptive Russian 1 Vocabulary Size Test

Band	N of Tests	Alpha	N of Items
1	25	0.93	30
2	141	0.92	30
3	127	0.93	30
4	123	0.93	30
5	98	0.94	30
Total	198		

Table R-3 shows that the internal consistency of each band was above 0.9, i. e., very good. To examine the goodness-of-fit of each individual item, the following statistics on the relationship between each individual item and all items of a band were calculated: the scale mean if the item was deleted; the corrected item-total correlation; and Cronbach's alpha if the item was deleted. Items below 0.3 in the column ***Corrected Item-Total Correlation*** do not correlate well with the overall score and, therefore, provide cause for concern (Field, 2018: 605). Items above the overall Cronbach's alpha of each band in the column ***Cronbach's Alpha if Item Deleted*** are also problematic, because if their removal raises alpha, then they are less reliable

than the average item (Field, 2018: 605). Tables R-4 to R-8 show the summary item statistics for each band of the receptive Russian 1 VST. Cells with misfitting values are set in bold and red.

Table R-4: Summary Item Statistics for the Receptive Russian 1 VST Band 1

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@1a	26.33	0.33	0.92
@1b	26.34	0.55	0.92
@1c	26.63	0.31	0.92
@2a	26.41	0.70	0.91
@2b	26.36	0.62	0.92
@2c	26.40	0.62	0.92
@3a	26.32	0.35	0.92
@3b	26.39	0.72	0.91
@3c	26.32	0.35	0.92
@4a	26.34	0.54	0.92
@4b	26.35	0.31	0.92
@4c	26.38	0.73	0.92
@5a	26.34	**0.16**	0.92

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@5b	26.36	0.35	0.92
@5c	26.33	**0.11**	0.92
@6a	26.51	0.65	0.92
@6b	26.36	0.64	0.92
@6c	26.55	0.62	0.92
@7a	26.35	0.63	0.92
@7b	26.38	0.51	0.92
@7c	26.46	0.55	0.92
@8a	26.64	0.43	0.92
@8b	26.44	0.61	0.92
@8c	26.50	0.73	0.91
@9a	26.40	0.55	0.92
@9b	26.43	0.61	0.92
@9c	26.41	0.62	0.92
@10a	26.38	0.53	0.92
@10b	26.38	0.70	0.92
@10c	26.37	0.62	0.92

Table R-5: Summary Item Statistics for the Receptive Russian 1 VST Band 2

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@11a	25.82	0.54	0.91
@11b	25.79	0.64	0.91
@11c	25.75	**0.29**	0.92
@12a	25.94	0.52	0.91
@12b	25.82	0.47	0.91
@12c	26.05	0.44	0.92
@13a	25.77	0.70	0.91
@13b	25.85	0.68	0.91
@13c	25.75	0.60	0.91
@14a	25.72	0.42	0.92
@14b	25.72	0.49	0.91
@14c	25.77	0.53	0.91
@15a	25.90	0.69	0.91
@15b	25.85	0.64	0.91
@15c	25.72	0.42	0.92
@16a	25.91	0.54	0.91
@16b	25.82	0.60	0.91

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@16c	25.74	0.46	0.91
@17a	26.01	0.62	0.91
@17b	26.04	0.41	0.92
@17c	25.74	0.52	0.91
@18a	25.87	0.64	0.91
@18b	25.89	0.50	0.91
@18c	25.75	0.50	0.91
@19a	25.74	0.31	0.92
@19b	25.77	0.48	0.91
@19c	25.76	0.57	0.91
@20a	25.72	0.47	0.91
@20b	25.76	0.45	0.91
@20c	25.81	0.51	0.91

Table R-6: Summary Item Statistics for the Receptive Russian 1 VST Band 3

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@21a	24.53	**0.00**	0.93
@21b	24.72	0.69	0.93

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@21c	24.78	0.48	0.93
@22a	24.63	0.62	0.93
@22b	24.59	**0.27**	0.93
@22c	24.69	0.60	0.93
@23a	24.80	0.59	0.93
@23b	24.57	0.45	0.93
@23c	24.65	0.66	0.93
@24a	24.55	0.38	0.93
@24b	24.56	0.48	0.93
@24c	24.66	0.62	0.93
@25a	24.65	0.53	0.93
@25b	24.88	0.45	0.93
@25c	24.63	0.62	0.93
@26a	24.68	0.71	0.93
@26b	24.75	0.65	0.93
@26c	24.91	0.65	0.93
@27a	24.56	0.38	0.93
@27b	24.72	0.69	0.93

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@27c	24.71	0.63	0.93
@28a	24.72	0.69	0.93
@28b	24.70	0.73	0.93
@28c	24.75	0.57	0.93
@29a	24.71	0.56	0.93
@29b	24.64	0.47	0.93
@29c	24.59	0.48	0.93
@30a	24.65	0.50	0.93
@30b	24.68	0.41	0.93
@30c	24.66	0.60	0.93

Table R-7: Summary Item Statistics for the Receptive Russian 1 VST Band 4

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@31a	24.54	0.39	0.93
@31b	24.56	0.42	0.93
@31c	24.46	0.53	0.93
@32a	24.47	0.68	0.92
@32b	24.43	0.52	0.93

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@32c	24.50	0.60	0.92
@33a	24.53	0.67	0.92
@33b	24.74	0.57	0.92
@33c	24.42	0.47	0.93
@34a	24.64	0.61	0.92
@34b	24.49	0.58	0.92
@34c	24.51	0.52	0.93
@35a	24.41	0.32	0.93
@35b	24.60	0.62	0.92
@35c	24.41	0.37	0.93
@36a	24.50	0.62	0.92
@36b	24.48	0.61	0.92
@36c	24.47	0.45	0.93
@37a	24.53	0.55	0.92
@37b	24.76	0.59	0.92
@37c	24.97	0.50	0.93
@38a	24.64	0.65	0.92
@38b	24.52	0.68	0.92

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@38c	24.59	0.48	0.93
@39a	24.67	0.63	0.92
@39b	24.54	0.58	0.92
@39c	24.46	0.45	0.93
@40a	24.47	0.46	0.93
@40b	24.63	0.48	0.93
@40c	24.61	0.59	0.92

Table R-8: Summary Item Statistics for the Receptive Russian 1 VST Band 5

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@41a	24.70	0.57	0.94
@41b	24.61	0.49	0.94
@41c	24.61	0.68	0.93
@42a	24.56	0.36	0.94
@42b	24.60	0.44	0.94
@42c	24.68	0.70	0.93
@43a	24.58	**0.26**	0.94
@43b	24.84	0.69	0.93

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@43c	24.90	0.42	0.94
@44a	24.73	0.59	0.94
@44b	24.71	0.71	0.93
@44c	24.62	0.62	0.94
@45a	24.66	0.64	0.93
@45b	24.62	0.61	0.94
@45c	24.78	0.51	0.94
@46a	24.63	0.64	0.93
@46b	24.94	0.67	0.93
@46c	24.81	0.67	0.93
@47a	24.67	0.56	0.94
@47b	24.67	0.67	0.93
@47c	24.70	0.75	0.93
@48a	24.70	0.63	0.93
@48b	24.78	0.68	0.93
@48c	24.96	0.56	0.94
@49a	24.79	0.66	0.93
@49b	24.63	0.45	0.94

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@49c	24.72	0.64	0.93
@50a	24.58	0.47	0.94
@50b	24.58	**0.19**	0.94
@50c	24.58	0.35	0.94

Tables R-4 to R-8 show that there were seven items in the receptive Russian 1 Vocabulary Size Test with conspicuous values. They will be inspected and revised, if needed, for version 2 of the receptive Russian 1 VST. Seven items out of 150, however, do not threaten the validity or reliability of the present version of the receptive Russian 1 VST. The data analyzed in this section, therefore, provide strong evidence that the present version (Russian 1.1) of the receptive Russian 1 VST exhibits excellent psychometric properties, meeting all professional standards of reliability and validity required in high-stakes testing.

Spanish: Receptive and Productive Vocabulary Size Tests 1

The Spanish 1 Vocabulary Size Test (VST) was modeled after the English Vocabulary Levels Test pioneered by Paul Nation (Nation, 1990). It uses the word frequency list of the Routledge Frequency Dictionary of Spanish (Davies & Davies, 2017). The VST measures how many of its words are known. In this chapter, evidence of validity and reliability of both, the receptive and productive Spanish 1 VST is presented. When test takers take the VST, their results are stored anonymously, without collecting any personal or technical information, to improve the VST. In the present report, data collected between April 2019 and March 2021 were analyzed to examine the overall validity and reliability of the receptive and productive Spanish 1 VST and to identify poorly performing items to be revised.

The Receptive Spanish 1 Vocabulary Size Test

The receptive Spanish 1 VST consists of five bands: the most frequent 1,000; 1,001 to 2,000; 2,001 to 3,000; 3,001 to 4,000; and 4,001 to 5,000 words. It includes ten clusters of six words each for each of the five bands. Each band is thus represented by 60 words. These words involve 30 nouns, 18 verbs, and 12 adjectives and are chosen at random from the 1,000 words of a band. Each cluster focuses on one part of speech, e. g., noun.

Each cluster contains six words and three synonyms, paraphrases, or gapped sentences (targets). Three of the six words are keys, i. e., they correspond to the three targets, while three words are additional distractors. For each target, the same six words are presented as multiple-choice options, one of which needs to be selected for each target. Each band, accordingly, consists of 30 items (targets). The maximum score per band is 30, i. e., 3 points per cluster. The maximum composite score for all five bands is 150, i. e., five times 30.

Correct responses were coded as 1 and incorrect responses as 0. Items that were not attempted were left blank. The maximum time allowed for the five-band test is 30 minutes. Tests that were completed in less than five minutes were removed to reduce the number of test takers who were responding only to a few items. Table S-1 shows the descriptive statistics of the receptive Spanish 1 VST.

Table S-1: Descriptive Statistics of the Receptive Spanish 1 Vocabulary Size Test

N	Mean	SE of Mean	Median	Std. Dev.	Minimum	Maximum
684	96.35	1.62	111	42.31	0	150

Table S-1 shows that there were 684 test takers. Total scores ranged from 0 to 150, covering the complete breadth of scores. The mean and the median were close to or within the upper third of the score range, indicating that there was a large number of test takers with high receptive vocabulary sizes.

To examine the overall reliability of the receptive Spanish 1 VST, Cronbach's alpha between the five bands of the receptive test was calculated. Cronbach's alpha is a measure of consistency, i. e., how consistent the results of all bands are to each other. It is commonly used as a measure of interrater reliability. Because it is a measure of internal consistency, it may be considered a measure of (internal) validity, i. e., it assesses how well different item sets measure the construct. If alpha is high, it may be assumed that all items measure the same construct, in this case receptive vocabulary size. Cronbach's alpha above 0.7 is considered acceptable, above 0.8, it is considered to be good, and above 0.9 very good. Table S-2 shows the number of tests administered, Cronbach's alpha, and the number of items, in this case, bands.

Table S-2: Cronbach's Alpha as a Measure of the Validity and Reliability of the Receptive Spanish 1 Vocabulary Size Test

N of Tests	Cronbach's Alpha	N of Items
531	0.94	5

Table S-2 shows that the reliability and internal validity of the receptive Spanish 1 VST is above 0.9 (very good), which supports the claim that it is highly valid and reliable. Note that the number of tests is different from the number of tests in Table S-1, because to calculate alpha across all bands, all bands need to have values. Some test takers only attempted one or more bands but not all five. For these test takers, Cronbach's alpha of all five bands could not be calculated.

To examine the internal consistency of each band, Cronbach's alpha was calculated for each band of 1,000 words. Each band consists of 30 items. Table S-3 shows the number of test takers, Cronbach's alpha, and the number of items for each band of the receptive Spanish 1 VST.

Table S-3: Cronbach's Alpha for Each Band of the Receptive Spanish 1 Vocabulary Size Test

Band	N of Tests	Alpha	N of Items
1	530	0.91	30
2	492	0.92	30
3	454	0.91	30
4	419	0.92	30
5	390	0.90	30
Total	684		

Table S-3 shows that the internal consistency of each band was at or above 0.9, i. e., very good. To examine the goodness-of-fit of each individual item, the following statistics on the relationship between each individual item and all items of a band were calculated: the scale mean if the item was deleted; the corrected item-total correlation; and Cronbach's alpha if the item was deleted. Items below 0.3 in the column ***Corrected Item-Total Correlation*** do not correlate well with the overall score and, therefore, provide cause for concern (Field, 2018: 605). Items above the overall Cronbach's alpha of each band in

the column ***Cronbach's Alpha if Item Deleted*** are also problematic, because if their removal raises alpha, then they are less reliable than the average item (Field, 2018: 605). Tables S-4 to S-8 show the summary item statistics for each band of the receptive Spanish 1 VST. Cells with misfitting values are set in bold and red.

Table S-4: Summary Item Statistics for the Receptive Spanish 1 VST Band 1

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@1a	24.56	0.37	0.91
@1b	24.51	0.42	0.91
@1c	24.58	0.57	0.91
@2a	24.55	0.47	0.91
@2b	24.49	0.36	0.91
@2c	24.58	0.65	0.90
@3a	24.51	0.42	0.91
@3b	24.55	0.48	0.91
@3c	24.72	0.44	0.91
@4a	24.55	0.48	0.91
@4b	25.13	**0.21**	0.91
@4c	24.68	0.59	0.90

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@5a	24.55	0.53	0.91
@5b	24.71	0.50	0.91
@5c	24.97	0.30	0.91
@6a	24.63	0.60	0.90
@6b	24.74	0.61	0.90
@6c	24.54	0.44	0.91
@7a	24.72	0.52	0.91
@7b	24.60	0.62	0.90
@7c	24.66	0.58	0.90
@8a	24.73	0.63	0.90
@8b	24.54	0.56	0.91
@8c	24.72	0.59	0.90
@9a	24.60	0.61	0.90
@9b	24.58	0.59	0.91
@9c	24.51	0.37	0.91
@10a	24.55	0.45	0.91
@10b	24.62	0.50	0.91
@10c	24.51	0.48	0.91

Table S-5: Summary Item Statistics for the Receptive Spanish 1 VST Band 2

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@11a	24.22	0.51	0.92
@11b	24.13	0.47	0.92
@11c	24.23	0.59	0.92
@12a	24.12	0.37	0.92
@12b	24.28	0.42	0.92
@12c	24.18	0.51	0.92
@13a	24.17	0.63	0.92
@13b	24.44	0.32	0.92
@13c	24.25	0.65	0.92
@14a	24.19	0.58	0.92
@14b	24.12	0.48	0.92
@14c	24.21	0.62	0.92
@15a	24.21	0.49	0.92
@15b	24.14	0.48	0.92
@15c	24.21	0.65	0.92
@16a	24.26	0.46	0.92
@16b	24.27	0.66	0.92

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@16c	24.40	0.58	0.92
@17a	24.21	0.51	0.92
@17b	24.20	0.60	0.92
@17c	24.18	0.56	0.92
@18a	24.26	0.49	0.92
@18b	24.38	0.57	0.92
@18c	24.28	0.57	0.92
@19a	24.45	0.53	0.92
@19b	24.24	0.60	0.92
@19c	24.60	0.35	0.92
@20a	24.40	0.55	0.92
@20b	24.40	0.55	0.92
@20c	24.21	0.66	0.92

Table S-6: Summary Item Statistics for the Receptive Spanish 1 VST Band 3

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@21a	23.47	0.58	0.91
@21b	23.54	0.42	0.91

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@21c	23.25	0.44	0.91
@22a	23.44	0.61	0.91
@22b	23.34	0.46	0.91
@22c	23.25	0.44	0.91
@23a	23.36	0.54	0.91
@23b	23.47	0.59	0.91
@23c	23.29	0.63	0.91
@24a	23.60	0.52	0.91
@24b	23.32	0.42	0.91
@24c	23.32	0.49	0.91
@25a	23.30	0.56	0.91
@25b	23.34	0.58	0.91
@25c	23.28	0.43	0.91
@26a	23.43	0.46	0.91
@26b	23.27	0.49	0.91
@26c	23.44	0.58	0.91
@27a	23.63	0.50	0.91
@27b	23.55	0.45	0.91

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@27c	23.49	0.60	0.91
@28a	23.43	0.61	0.91
@28b	23.55	**0.25**	0.91
@28c	23.43	0.57	0.91
@29a	23.30	0.44	0.91
@29b	23.53	0.61	0.91
@29c	23.27	0.47	0.91
@30a	23.52	0.36	0.91
@30b	23.46	0.37	0.91
@30c	23.32	0.54	0.91

Table S-7: Summary Item Statistics for the Receptive Spanish 1 VST Band 4

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@31a	23.29	0.61	0.92
@31b	23.29	0.56	0.92
@31c	23.32	0.52	0.92
@32a	23.21	0.47	0.92
@32b	23.31	0.56	0.92

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@32c	23.54	0.47	0.92
@33a	23.63	0.48	0.92
@33b	23.44	0.57	0.92
@33c	23.26	0.44	0.92
@34a	23.28	0.64	0.92
@34b	23.27	0.48	0.92
@34c	23.26	0.68	0.92
@35a	23.39	0.37	0.92
@35b	23.35	0.62	0.92
@35c	23.23	0.55	0.92
@36a	23.24	0.61	0.92
@36b	23.25	0.51	0.92
@36c	23.42	0.55	0.92
@37a	23.41	0.53	0.92
@37b	23.49	0.56	0.92
@37c	23.45	0.62	0.92
@38a	23.28	0.63	0.92
@38b	23.75	**0.26**	0.93

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@38c	23.47	0.49	0.92
@39a	23.34	0.52	0.92
@39b	23.33	0.57	0.92
@39c	23.30	0.55	0.92
@40a	23.32	0.60	0.92
@40b	23.32	0.63	0.92
@40c	23.59	0.46	0.92

Table S-8: Summary Item Statistics for the Receptive Spanish 1 VST Band 5

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@41a	19.35	0.54	0.89
@41b	19.36	0.58	0.89
@41c	19.27	0.52	0.89
@42a	19.27	0.52	0.89
@42b	19.24	0.56	0.89
@42c	19.34	0.57	0.89
@43a	19.54	0.50	0.89
@43b	19.26	0.53	0.89

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@43c	19.66	0.33	0.90
@44a	19.27	0.47	0.89
@44b	19.25	0.54	0.89
@44c	19.43	0.46	0.89
@45a	19.25	0.52	0.89
@45b	19.44	0.54	0.89
@45c	19.39	0.50	0.89
@46a	19.56	0.42	0.89
@46b	19.53	**0.22**	0.90
@46c	19.72	**0.27**	0.90
@47a	19.23	0.56	0.89
@47b	19.21	0.55	0.89
@47c	19.27	0.62	0.89
@48a	19.57	0.46	0.89
@48b	19.52	0.44	0.89
@48c	19.36	0.49	0.89
@49a	19.21	0.49	0.89
@49b	19.67	**0.29**	0.90

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@49c	19.54	0.46	0.89
@50a	19.13	0.44	0.89
@50b	19.84	**0.21**	0.90
@50c	19.76	**0.18**	0.90

Tables 4-8 show that there were eight items in the receptive Spanish 1 Vocabulary Size Test with conspicuous values. They will be inspected and revised, if needed, for version 2 of the receptive Spanish 1 VST. Eight items out of 150, however, do not threaten the validity or reliability of the present version of the receptive Spanish 1 VST. The data analyzed in this section, therefore, provide strong evidence that the present version (Spanish 1.1) of the receptive Spanish 1 VST exhibits excellent psychometric properties, meeting all professional standards of reliability and validity required in high-stakes testing.

The Productive Spanish 1 Vocabulary Size Test

The productive vocabulary size test consists of 18 sentences, one for each targeted word, which is partially missing. Each band is thus represented by 18 words. These words involve 9 nouns, 6 verbs, and 3 adjectives chosen at random from the 1,000 words of a band. The maximum score per band is 18. The maximum composite score for all five bands is 90, i. e., five times 18.

The targeted words appear towards the end of the sentence to establish their meaning. The first few letters are given to disambiguate the word from other possible words. As many letters are provided as needed to disambiguate any given word, up to 50 % of the letters of the word, i. e., if a word consists of an odd number of letters, a maximum of half of the letters minus one are provided. All words of a particular sentence are part of the same band or a more frequent band. In Bands 1 and 2, the partially missing words are uninflected: verbs, for example, appear in their infinitive form. In Bands 3, 4, and 5, partially missing words may be inflected, i. e., grammatical knowledge may be required. Words are scored correct only if they are 100 % correct, including orthography and grammar.

Correct responses were coded as 1 and incorrect responses as 0. Items that were not attempted were left blank. Table S-9 shows the descriptive statistics of the productive Spanish 1 VST. Tests that were completed in less than 5 minutes were removed.

Table S-9: Descriptive Statistics of the Productive Spanish 1 Vocabulary Size Test

N	Mean	SE of Mean	Median	Std. Dev.	Minimum	Maximum
367	40.25	0.99	43	19.04	0	84

Table S-9 shows that there were 367 test takers. Total scores ranged from 0 to 84, covering most of the score range. The mean and the median were slightly below the midpoint of the

scale (45) with a standard deviation of close to 20, indicating that there was a broad cross section of vocabulary sizes.

To examine the overall reliability of the productive Spanish 1 VST, Cronbach's alpha between the five bands of the productive test was calculated. Cronbach's alpha is a measure of consistency, i. e., how consistent the results of all bands are to each other. Because it is a measure of internal consistency, it may be considered a measure of (internal) validity, i. e., it assesses how well different item sets measure the construct. If alpha is high, it may be assumed that all items measure the same construct, in this case productive vocabulary size. Cronbach's alpha above 0.7 is considered acceptable, above 0.8, it is considered to be good, and above 0.9 very good. Table S-10 shows the number of tests administered, Cronbach's alpha, and the number of items, in this case, bands.

Table S-10: Cronbach's Alpha as a Measure of the Validity and Reliability of the Productive Spanish 1 Vocabulary Size Test

N of Tests	Cronbach's Alpha	N of Items
327	0.93	5

Table S-10 shows that the reliability and internal validity of the productive Spanish 1 VST was above 0.9 (very good), which supports the claim that it is highly valid and reliable. Note that the number of tests is different from the number of tests in Table S-9, because to calculate alpha across all bands, all bands need to have values. Some test takers only attempted one or

more bands but not all five. For these test takers, Cronbach's alpha of all five bands could not be calculated.

To examine the internal consistency of each band, Cronbach's alpha was calculated for each band of 1,000 words. Each band consists of 18 items. Table S-11 shows the number of test takers, Cronbach's alpha, and the number of items for each band of the productive Spanish 1 VST.

Table S-11: Cronbach's Alpha for Each Band of the Productive Spanish 1 Vocabulary Size Test

Band	N of Tests	Alpha	N of Items
1	181	0.73	18
2	131	0.78	18
3	144	0.74	18
4	59	0.84	18
5	54	0.89	18
Total	367		

Table S-11 shows that the internal consistency of Bands 1 and 2 was above 0.7 (acceptable) and that it was good for Bands 2, 4, and 5 (close to or above 0.8). To examine the goodness-of-fit of each individual item, the following statistics on the relationship between each individual item and all items of a band were calculated: the scale mean if the item was deleted; the corrected item-total correlation; and Cronbach's alpha if the item was deleted. Items below 0.3 in the column ***Corrected***

Item-Total Correlation do not correlate well with the overall score and may, therefore, provide cause for concern (Field, 2018: 605). Items above the overall Cronbach's alpha of each band in the column *Cronbach's Alpha if Item Deleted* are also problematic, because if their removal raises alpha, then they are less reliable than the average item (Field, 2018: 605). Tables S-12 to S-16 show the summary item statistics for each band of the productive Spanish 1 VST. Cells with misfitting values are set in bold and red.

Table S-12: Summary Item Statistics for the Productive Spanish 1 VST Band 1

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@1	12.18	**0.27**	0.72
@2	12.50	**0.22**	0.73
@3	12.07	0.45	0.71
@4	12.13	0.33	0.72
@5	12.27	**0.22**	0.73
@6	12.19	**0.24**	0.72
@7	12.08	**0.25**	0.72
@8	12.15	0.34	0.71
@9	12.14	0.38	0.71
@10	12.06	0.43	0.71

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@11	12.00	0.35	0.72
@12	12.32	**0.21**	0.73
@13	12.24	0.40	0.71
@14	12.27	**0.18**	0.73
@15	12.11	0.34	0.72
@16	12.77	0.30	0.72
@17	12.08	0.41	0.71
@18	12.59	0.42	0.71

Table S-13: Summary Item Statistics for the Productive Spanish 1 VST Band 2

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@19	12.42	**0.25**	0.78
@20	11.96	0.38	0.77
@21	11.96	**0.29**	0.78
@22	11.98	0.43	0.77
@23	12.24	0.47	0.77
@24	11.93	0.45	0.77
@25	12.04	0.38	0.77

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@26	12.27	0.39	0.77
@27	11.93	0.46	0.77
@28	11.95	0.49	0.77
@29	12.41	0.31	0.78
@30	12.37	0.44	0.77
@31	11.86	0.59	0.77
@32	11.89	0.42	0.77
@33	12.07	0.33	0.78
@34	12.63	**0.29**	0.78
@35	12.11	0.31	0.78
@36	12.09	**0.24**	0.78

Table S-14: Summary Item Statistics for the Productive Spanish 1 VST Band 3

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@37	10.64	0.38	0.73
@38	10.84	0.38	0.73
@39	10.98	0.33	0.73
@40	10.65	0.39	0.73

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@41	10.84	0.32	0.73
@42	10.69	0.36	0.73
@43	11.12	**0.14**	0.75
@44	10.68	0.30	0.73
@45	10.75	0.34	0.73
@46	11.47	**0.24**	0.74
@47	10.69	0.47	0.72
@48	11.39	0.34	0.73
@49	10.92	0.41	0.72
@50	10.88	0.40	0.72
@51	11.26	0.32	0.73
@52	11.35	0.33	0.73
@53	10.98	0.38	0.73
@54	11.15	**0.24**	0.74

Table S-15: Summary Item Statistics for the Productive Spanish 1 VST Band 4

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@55	8.54	**0.26**	0.84

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@56	8.42	0.39	0.83
@57	9.15	0.43	0.83
@58	8.81	0.47	0.83
@59	8.68	0.50	0.83
@60	8.80	0.51	0.83
@61	8.39	**0.25**	0.84
@62	9.12	0.42	0.83
@63	8.59	0.40	0.83
@64	9.15	0.49	0.83
@65	8.58	0.52	0.83
@66	8.53	0.44	0.83
@67	8.64	0.59	0.82
@68	8.76	0.33	0.84
@69	9.19	**0.24**	0.84
@70	9.05	0.61	0.82
@71	8.83	0.47	0.83
@72	8.66	0.46	0.83

Table S-16: Summary Item Statistics for the Productive Spanish 1 VST Band 5

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@73	8.98	0.62	0.88
@74	8.94	0.40	0.89
@75	8.80	0.73	0.88
@76	8.70	0.49	0.89
@77	8.87	0.45	0.89
@78	8.74	0.74	0.88
@79	8.94	0.65	0.88
@80	9.09	0.58	0.89
@81	9.22	0.43	0.89
@82	9.46	**-0.18**	**0.90**
@83	9.09	0.54	0.89
@84	9.22	0.53	0.89
@85	8.98	0.50	0.89
@86	8.80	0.66	0.88
@87	9.04	0.44	0.89
@88	8.80	0.46	0.89
@89	8.65	0.45	0.89

	Scale Mean if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
@90	8.85	0.70	0.88

Tables S-12 to S-16 show that there were 18 items with conspicuous values in the productive Spanish 1 VST. Most of them were in Band 1. They will be inspected and revised, where needed, for version 2 of the productive Spanish 1 VST. The probability that many, if not most, of these items will not exhibit any issues is relatively high. Meanwhile, the data analyzed in this section provide sufficient evidence that the present version (Version 1.1) of the productive Spanish 1 VST exhibits sound psychometric properties, meeting all professional standards of reliability and validity required in high-stakes testing.

References

Buckwalter, T., & Parkinson D. (2011). A Frequency Dictionary of Arabic: Core Vocabulary for Learners. London: Routledge.

Davies, M., & Davies, K. H. (2017). A Frequency Dictionary of Spanish: Core Vocabulary for Learners. London: Routledge.

De Mauro, T., Mancini, F., Vedovelli, M., & Voghera, M. (1993). Lessico di frequenza dell'italiano parlato. Milano: Etaslibri. Online: http://badip.uni-graz.at/ (September 27, 2021)

De Mauro, T. & Moroni, G. G. (1996). DIB: Dizionario di base della lingua Italiana. Torino: Paravia.

Field, A. (2018). Discovering Statistics Using IBM SPSS Statistics. North American Edition. Los Angeles and others: Sage.

Hacking, J., Rubio, F., & Tschirner, E. (2019). Vocabulary Size, Reading Proficiency and Curricular Design: The Case of College Chinese, Russian and Spanish. In: S. Gass & P. Winke (eds.), Foreign Language Proficiency in Higher Education (Educational Linguistics 37) (pp. 25-44). Cham: Springer Nature Switzerland.

Huhta, A., Alderson, J. C., Nieminen, L., & Ullakonoja, R. (2011). Diagnosing Reading in L2: Predictors and Vocabulary Profiles. Paper presented at the ACTFL CEFR Conference, Provo, UT, August 4, 2011.

Institute for Test Research and Test Development (n.d.) Vocabulary Tests. Online: https://itt-leipzig.de/about-the-vocabulary-tests-2/?lang=en (September 27, 2021).

Jones, R. & Tschirner, E. (2006). Frequency Dictionary of German: Core Vocabulary for Learners. London: Routledge.

Laufer, B., & Nation, I. S. P. (2012). Vocabulary. In: S. Gass & A. Mackey (eds.), The Routledge Handbook of Second Language Acquisition (pp. 163-176). New York: Routledge.

Leech, G., Rayson, P., & Wilson, A. (2001). Word Frequencies in Written and Spoken English: Based on the British National Corpus. London: Longman.

Lonsdale, D., & Le Bras, Y. (2009). A Frequency Dictionary of French: Core Vocabulary for Learners. London: Routledge.

Milton, J. (2009). Measuring Second Language Vocabulary Acquisition. Bristol: Multilingual Matters.

Milton, J. (2010): The Development of Vocabulary Breadth across the CEFR Levels. In: I. Bartning, M. Martin, & I. Vedder (eds.), Communicative Proficiency and Linguistic Development: Intersections between SLA and Language Testing Research (EuroSLA Monographs Vol. 1) (pp. 211–232). Online: http://www.eurosla.org/monographs/EM01/EM01home.php (May 26, 2021).

Nation, I.S.P. (1990). Teaching and Learning Vocabulary. New York: Newbury House.

Qian, D. D., & Lin, L. H. (2019). Vocabulary Knowledge and Language Proficiency. In: S. Webb (ed.), The Routledge Handbook of Vocabulary Studies (pp. 58–77). New York: Routledge.

Routledge (n.d.). Routledge Frequency Dictionaries. Online: https://www.routledge.com/Routledge-Frequency-Dictionaries/book-series/RFD (September 27, 2021).

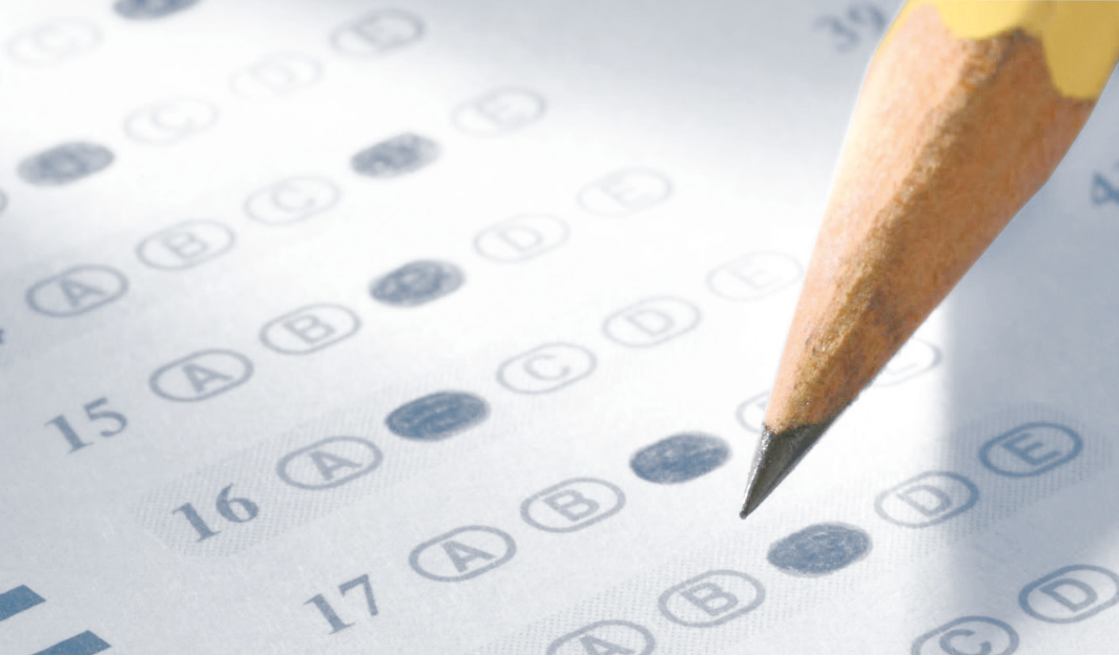
Schmitt, N. (2008). Instructed Second Language Vocabulary Learning. In: Language Teaching Tesearch, 12(3), (pp. 329–363).

Sharoff, S., Umanskaya, E., & Wilson, J. (2013). A Frequency Dictionary of Russian: Core Vocabulary for Learners. London: Routledge.

Stæhr, L. S. (2008). Vocabulary Size and the Skills of Listening, Reading and Writing. In: Language Learning Journal, 36(2), (pp. 139–152).

Tschirner, E. (2019). Der rezeptive Wortschatzbedarf im Deutschen als Fremdsprache. In: E. Peyer, T. Studer, & I. Thonhauser (eds.), IDT 2017, Band 1: Hauptvorträge (pp. 98–111). Berlin: Erich Schmidt.

Xiao, R., Rayson, P., & McEnery, T. (2009). A Frequency Dictionary of Mandarin Chinese: Core Vocabulary for Learners. London: Routledge.



INSTITUTE FOR TEST RESEARCH
AND TEST DEVELOPMENT