Vocabulary Learning and Instruction
Volume 11, Issue 1, December 2022
http://vli-journal.org

A Proposed Taxonomy of Test-Taking Action and Item Format in Written Receptive Vocabulary Testing

Jeffrey Martin
Momoyama Gakuin University

Abstract

The functioning of a vocabulary testing instrument rests in part on the test-taking actions made possible for examinees by item format, an aspect of test development that warrants consideration in second-language vocabulary research. For example, although iterations of the written receptive vocabulary levels test (VLT) have integrated improvements in lexis sampling and distractor-item creation (i.e., Beglar & Hunt, 1999; Nation, 1983, 1990; Schmitt et al., 2001; Webb et al., 2017), its clustered form-meaning matching format has remained fundamentally unchanged. This study qualitatively explores the influence of this test item format on test-taking actions observed when taking the updated vocabulary levels test (UVLT, Webb et al, 2017). Data from a think-aloud protocol and retrospective interviewing indicated the predominant use of test-taking strategies for answering test items on the UVLT, such as bidirectional matching and elimination of cluster options, and that these actions enabled correct responses for clusters of target vocabulary about which the test taker demonstrated partial or even no knowledge. This evidence at the interface of test taker and test draws attention to the interconnection of estimating learners' vocabulary knowledge and the action possibilities provided by item format on vocabulary tests. Such affordances are hierarchically structured in a proposed *Taxonomy of Test-taking Actions Afforded by Receptive Vocabulary Test Format* as a heuristic to evaluate the influences of test format on written receptive vocabulary assessment.

Keywords: receptive vocabulary knowledge, test modality, think-aloud protocol, affordance, heuristic

1 Introduction

Item format has become a salient aspect of second-language (L2) vocabulary testing due to a range of empirical research demonstrating how test format appears to tap different aspects of L2 vocabulary knowledge (VK) (e.g., Kremmel & Schmitt, 2016; Laufer & Goldstein, 2004; McLean et al., 2020). Schmitt et al. (2020) highlighted the need for rigorous vocabulary testing for specified purposes and that item format is one essential factor. The wide variety of vocabulary testing that is now commonly available includes the recalling and recognizing of lexis, the eliciting of lexical form and lexical meaning, and the selecting and matching of

Please cite this article as: Martin, J. (2022). A proposed taxonomy of test-taking action and item format in written receptive vocabulary testing. *Vocabulary Learning and Instruction, 11*(1), 1–16. https://doi.org/10.7820/vli.v11.1.martin

options to solve for test items. Differences in testing outcomes bear implications about what can be inferred from test scores regarding the VK that L2 learners can employ within given settings of L2 communication.

To further explore this issue, direct observation of test-taking behavior can help to clarify what test-taking actions are enabled within an item format and to then make inferences about test-taking actions enabled by other item formats. This study centers on the observation of an assessment performed on a vocabulary test featuring an item-clustered form-meaning matching format. In this article, I begin by outlining the importance of item format on test scores. I also outline the importance of directly observing the actions taken on an item-clustered format vocabulary test, as well as a review of the procedure. I then present a set of actions observed to be made possible within this item format and the variety of these actions that were realized in answering different clusters. Finally, based on the analysis of these observed patterns, I propose a taxonomy of test-taking actions by test item format that hierarchically categorizes a range of target and non-target VK argued to be usable to solve for given item types.

2 Background

Vocabulary development is a multifaceted process for L2 learners (Nation, 2020; Read, 2020; Schmitt et al., 2020). Research suggests that VK elicited by vocabulary tests is partial and dependent on test item format (Kremmel & Schmitt, 2016; Laufer & Goldstein, 2004; McLean et al., 2020) and, therefore, may not be uniformly employable across distinct kinds of language performance (i.e., VK that facilitates reading a book versus writing a letter). To best meet the needs of pedagogic and research purposes, "careful thought needs to be given to the item type that is used to make sure that it is suited to the kind of knowledge it is supposed to measure" (Nation & Webb, 2011, p. 219). Thus, two important considerations for selecting or developing a vocabulary test are that the instrument can (a) reasonably target intended VK in test takers and (b) sufficiently estimate the presence of that knowledge.

Although starting from an exploratory stance, the following range of format types became relevant to this study. One common vocabulary modality of item formatting is recognition testing, where test takers recognize and choose a correct option for a test item listed among a set of distractor items. Another common test modality is recall testing, where test takers recall information to correctly answer test items without answer options to choose from. Additionally, these two test modalities are commonly formatted to elicit the specific aspects of receptive VK (e.g., L2 word *meaning recognition*) and productive VK (e.g., L2 word *form recall*). For instance, Schmitt (2014) presented evidence that producing L2 form is a "deeper" level of vocabulary mastery than the comprehending of L2 word meaning encountered in L2 input. An additional item format is a multiple matching item type that involves matching L2 word forms and L2 word meanings within a cluster of target vocabulary and definitions.

This study centered on the functioning of the item-clustered form-meaning matching format of the updated vocabulary levels test (UVLT; Webb et al., 2017).

The test was designed to be "a measure of receptive VK indicating the degree to which test takers may be able to understand the meanings of words that they encounter in written text" (p. 57). It was stated to not measure productive VK nor the influence of word frequency on word difficulty. The UVLT is the most recent iteration of the vocabulary levels test (VLT). Its 150 target words are sourced from the top five frequency bands of Nation's (2012) British National Corpus (BNC)/Corpus of Contemporary American English (COCA) word family. Each band is represented by 10 clusters, and each cluster features three target words. The UVLT is preceded by Nation (1983, 1990), Beglar and Hunt (1999), and Schmitt et al. (2001). All are commonly structured in item clusters where test takers match six words (three target words and three distractors) with three definitions.

Quantitative studies have brought insight into the functioning of the VLT. For example, Kamimoto (2014) made an experimental version of the VLT by Nation (1990) that combined sets of three clusters to make larger clusters of 18 words and 9 definitions. This was in order to compare the relative effect of cluster size on testing outcomes. He found a nearly 19% inflation in test scores from the original test compared to the experimental version. In practice, estimating the lexical challenge an L2 learner might encounter with a given reading text would be hampered by a 19% inflation of assessment when considering the narrow thresholds for L2 reading referenced in the literature (e.g., 98% of words in a text known to facilitate reading fluency [Nation, 2006]). Additionally, Ha (2022) found correlations between correctly answered word items within clusters of the UVLT. However, an itemized language test should maintain test item independence in order to ensure fairness. This is to avoid disproportionately favoring test takers who answer select items correctly, over others who do not, by preferentially assisting them in correctly answering subsequent items (Bond et al., 2020). Score inflation and item interdependence interfere with the meaningfulness of test scores.

Directly observing clustered-format vocabulary testing is also considered due to the outcomes of Rasch-based analysis (Rasch, 1960), as was used by Beglar and Hunt (1999), Schmitt et al. (2001), and Webb et al. (2017). Webb et al.'s (2017) preliminary validity evidence for the UVLT found predictable item fit figures within the routinely accepted threshold of two standard deviations of the standardized mean (Bond et al., 2020). Notably for person measures, however, the threshold for removing test takers giving highly unpredictable/misfitting responses was inclusive of up to five standard deviations (Rasch outfit cutoff $z > 5.0$; Webb et al., 2017, pp. 38–39). For items and persons, the Rasch model anticipates that higher ability test takers will more likely respond correctly to items from high to low difficulty, whereas lower ability test takers will more likely only respond correctly to relatively easier items. If the opposite regularly occurs within the data, it may suggest that the test is garnering unintended or unpredictable test-taking behavior.

Comparing Rasch item reliability figures between vocabulary tests may not signal substantial differences in unintended test-taking behavior due to ceiling effects (upper limit of 1.0). However, Rasch item separation values scale with no upper limit and are not subject to ceiling effects (Smith, 2001). They represent statistically separable levels of difficulty instantiated by item functioning. Responses not accounted for by the model can degrade item separation figures, which lessens

confidence in the test's replicability. Item separation is largely affected by the quality of the items, given a sufficient number of items comprise the test (Bond et al., 2020).

Webb et al. (2017) stated that the two forms of the UVLT (versions A and B) had reliability (and separation) estimates of 0.96 (4.72) and 0.96 (4.81), respectively. The UVLT is comparable to the vocabulary size test (VST; Beglar, 2010) if only the higher frequency target items are considered (e.g., a shortened version covering levels 1k – 4k). The VST is a written meaning-recognition test of 140 items that estimates the total size of a learner's vocabulary: ten words each per fourteen 1000-word bands of the BNC word list by Nation (2006). Beglar (2010) demonstrated consistent item reliability and item separation for the full test (1k – 14k, 0.96 and 5.22) by comparing it to a subset of the test's items. A version of only the first four frequency levels (1k – 4k, only 40 items) showed favorable item reliability and separation figures of 0.98 and 6.25. Compared to the UVLT (150 items, top five frequency levels), the separation figures indicate a 30% increase in the distinguishability of item difficulty for the shortened VST. The VST is a meaning-recognition test, and the UVLT is a clustered-matching test. The sampling of the VST and UVLT was comparable (BNC and BNC/COCA word lists). This evidence for differences in test functioning is limited but it invites a qualitative look into the actions enabled (afforded) by clustered vocabulary items.

Termed by Gibson (1979), affordances are what the environment "*offers* the animal, what it *provides* or *furnishes*, either for good or ill", and evolutionarily, "they are unique for that animal" (italics in the original; pp. 119–120). Gibson's theory of affordance in nature was applied by Norman (1988) to user interface design in human–computer interaction, where actionability with objects or computer screens becomes salient depending on product design and user perception. Norman's (1988) conceptualization of affordances includes the influence of cultural conventions in how humans perceive objects, which are subject to the physical constraints of the objective existence of affordances in the given environment. This conceptualization is taken to be applicable to actions and strategies afforded by item format as a test taker interfaces with a paper-based vocabulary test.

Affordance is also conceptualized within the sociocognitive approach to SLA (van Lier, 2004). For example, Atkinson et al. (2018) studied the activity of collaborative baking, "a form of *triadic interaction*, wherein individuals focus shared attention and action on co-active environmental affordances in completing a task" (italics in the original; p. 477). Churchill et al. (2010) conducted a sociocognitive study of the interaction between a tutor and a student working together on a grammar worksheet. These studies investigated affordances within social interaction. In contrast, the current study is about one participant's engagement with a vocabulary test, so affordance is bound to Norman's conceptualization (1988).

The primary source of data for analysis in this study was a think-aloud protocol (Johnstone et al., 2006), where test takers verbalize their thought processes as they complete a task. Wilson's (1994) paper on the completeness of data retrievable by a think-aloud protocol highlighted the inability to elicit unconscious thoughts, nor all conscious thoughts, of participants, but that its concurrent verbal

reporting held advantages of accuracy and richness over retrospective questioning. Topic is also important because during a think-aloud protocol, "self-presentational concerns are more likely to be operative in […] social domains" than in problem-solving domains due to social topics being potentially more sensitive (p. 251). The problem-solving task of completing a vocabulary test in this study seems appropriate for a think-aloud protocol. As a complimentary data source, interview data can also attest to a test taker's thoughts and actions (Schmitt, 1999; Schmitt et al., 2001).

An exploratory investigation of test-taker actions and testing outcomes holds no preconceptions. The notion of affordance in this study became relevant during the analysis stage as a way to categorize the actions that were observed to be made possible by the clustered item format. A resulting taxonomy represents a hierarchical structure of test-taking actions in relation to the written receptive vocabulary testing formats summarized above. The proposed taxonomy details relationships of affordances, and it may provide a heuristic that is generalizable within vocabulary testing research. It is open to be rebutted or corroborated in proportion to the weight of the evidence presented. The considerations detailed above drove the formulation of the research questions given as follows:

RQ1: What lexical knowledge does a test taker demonstrate on the UVLT?

RQ2: What test-taking actions are afforded to a test taker when taking the UVLT?

RQ3: How and to what extent can the data about vocabulary knowledge and test-taking actions lead to categorizations and inferences regarding the formatting of written receptive vocabulary testing?

3 Methodology

3.1 Participant

The participant was a 27-year-old Japanese female working at the patient service counter at a hospital in the Tokyo area. Naoko, a pseudonym, had never taken the UVLT or any prior VLTs. She reported earning a TOEIC score of 420 points during her second year at university. She also reported studying English since that time for social and travel reasons. Naoko and I met to discuss the study, which would entail three meetings over a 2-week period, and the compensation would include a gift card. She gave her informed consent to participate in the think-aloud protocol, to be interviewed, and to be audio recorded. All data and field notes were kept in a secure location offline.

3.2 Materials

The UVLT (Webb et al., 2017) is a 150-item test sampled from the five most frequent 1000-word family bands of the BNC/COCA family word list by Nation (2012). The two versions (A and B) of the UVLT were developed to be

It should be answered in the following way.

	game	island	mouth	movie	song	yard
land with water all around it		✓				
part of your body used for eating and talking			✓			
piece of music					✓	

Figure 1. Example of the cluster of items on the UVLT (Webb et al., 2017).

1 business	
2 clock	_____ part of a house
3 horse	_____ animal with four legs
4 pencil	_____ something used for writing
5 shoe	
6 wall	

Figure 2. Example of the cluster of items on the VLT (Schmitt et al., 2001).

equivalent. The UVLT Version B was used in this study and is henceforth referred to as the UVLT. Each 1000-word band is represented by a 30-word sample in ten clusters, each including three target words. Each band consists of five clusters of nouns, three clusters of verbs, and two clusters of adjectives. Each cluster contains three definitions in English and six vocabulary items (three target words and three distractors). An instructional example from the UVLT is shown in Figure 1. An example of the VLT (Schmitt et al., 2001) is also provided in Figure 2 to illustrate the continuity of the clustered matching format across iterations of the VLT. The clusters were numbered in this study to help illustrate findings, representing the *n*th cluster as ordered on the UVLT version B. The test is accessible at https://www.edu.uwo.ca/faculty-profiles/docs/other/webb/NVLT-VERSION-B.pdf.

3.3 Procedure

The sequence of Naoko's participation in the study is detailed in Table 1. First, Naoko completed the odd-numbered clusters, 25 of the 50 clusters, at her own pace. She was informed that the interview may cover her answers of the odd-numbered items. One week after completing the odd-numbered clusters, Naoko completed the even-numbered clusters while following a think-aloud protocol (Johnstone et al., 2006; Wilson, 1994), where she was asked to concurrently verbalize her thought processes as she completed the clusters of items. In cases where Naoko stopped talking, I provided follow-up prompts such as, "What are you thinking now?" and "Please say what you're thinking." Naoko's description of her decision-making process was transcribed. Her responses for all 150 items were tallied and reviewed. I then used these data to prepare for the retrospective interview.

The semi-structured interview was about how she engaged with the UVLT (revisiting clusters from both even- and odd-numbered clusters). She was free to participate in either English or Japanese. Examples of the interview questions for Naoko were "Please tell me what the word means?" and "How did you decide on that answer?". Next, I transcribed the interview session and organized all data

Table 1. Sequence of Naoko's participation

Meeting	Task	Setting
Day 1	Complete the odd-numbered clusters. Following this, self-reflect on the experience of taking the test	Independently completed, no time limit
Day 7	Complete the remaining even-numbered clusters following a think-aloud protocol	Carried out with the researcher, live recording of the think-aloud, no time limit
Day 14	Retrospective interview	With the researcher, live recording of the interaction, about 40 minutes

into spreadsheets for detailed coding and analysis. As a researcher interested in L2 vocabulary, I was aware of L2 vocabulary studies on partial word knowledge, the influence of cognates, and so on, but I held no preconceptions regarding Naoko's engagement with the UVLT, nor any preconceived goals to later organize observed actions by item format. The resulting coding system was an emergent outcome of analyzing the study's data. Coded transcripts for the think-aloud protocol and the interview are placed in online supplemental materials that are retrievable from the Open Science Framework (OSF; https://osf.io/ypxze/).

3.4 Analysis

Analysis of the transcribed data saw the emergence of patterns for coding using frameworks outlined by Saldaña (2016). Initially, the data were coded for attributes such as cluster number and frequency band. Next, coding for the evidence of VK and patterns of action emerged, and this led to the creation of a code book (presented in supplemental materials). The actions observed in Naoko's test taking suggested that she was making decisions in a hierarchy of patterns, with some leveled above others. Such data called for taxonomic coding, which Saldaña (2016) described as a way to discover the "knowledge that people use to organize their behaviors and interpret their experiences" (p. 157). The resulting taxonomic coding of the think-aloud data and interview data was revised over repeated analysis.

After coding the data, the data and codes were shared in a data session with seven colleagues who were familiar with the project and had experience in second-language studies. All together and in smaller groups, the members of this session analyzed the transcripts of the think-aloud protocol and the interviews. Their feedback confirmed much of the coding scheme and provided additional nuance to the coding and analysis.

4 Findings

4.1 Vocabulary Knowledge Demonstrated

Observed degrees of VK ranged from no VK to robust VK and observations varied by cluster. Partial word knowledge was evident in at least two ways: knowledge of word parts and knowledge from L1 loanword equivalents.

For example, the think-aloud data (lines 82–83) for cluster c38 showed that knowledge of the prefix "trans," meaning "to change," allowed Naoko to match the unknown word "transplant" to its meaning of "move something to another place." Another example, loanword knowledge, enabled Naoko to eliminate the distractor word 'tank' in cluster c14. In the interview, I asked if she knew that word, and she responded, "Like water tank?", in reference to one of its loanword equivalences in Japanese (Interview, lines 20–24). She did not describe additional meanings for tank, but her knowledge was sufficient to eliminate this distractor option.

Naoko spoke of having no VK for some target items during the think-aloud protocol session and the interview. For some of these items, she guessed incorrectly. For example, "Twenty-eight. Just my guess but 'exceed'. 'Goes beyond the limit' is 'decline'. 'Take in' is 'link'. I don't know these words. I don't know the meaning" (Think-aloud, line 50). Nevertheless, numerous other items were answered correctly despite Naoko's lack of VK for the target words. A prime example from the interview data was cluster c29, with the targets "approximate," "frequent," and "prior," where Naoko stated having no knowledge despite successfully matching all three of the cluster's form-meaning pairs:

> *The words are kinda like… yeah, I don't really know them, but I knew some of the left side (definitions). Like maybe 'happening often' is frequency, but I wasn't sure, ya know? But I didn't know 'approximate' and 'prior'… I didn't know the meanings at all. I know 'graphic and 'vital', but their meanings don't fit with the left side.* (Interview, lines 56–60)

For the remaining distractor word "pale," she commented that she knew the color. With minimal or no knowledge of the three target words in cluster c29, Naoko correctly matched them to their meanings merely by eliminating cluster options.

Analysis of these data made it apparent that partial VK aided Naoko in matching form and meaning. For cluster c26, Naoko states, "Agree. consent, enforce, exhibit, retain, specify, target. (2-second pause) Hmm. I don't know. Next, 'say clearly'… (2-second pause) I don't know anything" (think-aloud, line 44). From this segment of think-aloud data, it was not explicit whether she merely eliminated options, used unstated knowledge of the target words, or simply guessed, but she matched all items correctly for the cluster despite her stating that she had no knowledge. Naoko's solving for this cluster resembles patterns of her solving for other clusters and generally illustrates the success she could achieve by using degrees of VK.

4.2 Actions Afforded to a Test Taker When Taking the UVLT

The format of the UVLT appears to afford test takers the additional ability to switch matching direction between linking target forms to meaning and the list of meanings to target forms. Naoko was observed to overwhelmingly process meanings listed in the left column of each cluster first and then to match them to target word forms in the top row of the cluster. Having a definition in mind and then selecting a target form from a list of options appears to tap a test-taker's ability to recognize L2 word form, an aspect of word knowledge that is distinct from meaning recognition

Table 2. Observations of directions in target matching during think-aloud protocol

Observed matching direction	Cluster and frequency band				
From target form to target meaning			c28 (3k)	c32 (4k) c40 (4k)	c42 (5k)
From target meaning to target form	c2 (1k) c4 (1k) c6 (1k) c8 (1k) c10 (1k)	c12 (2k) c14 (2k) c16 (2k) c18 (2k) c20 (2k)	c22 (3k) c24 (3k) c26 (3k) c30 (3k)	c34 (4k) c36 (4k) c38 (4k)	c44 (5k) c46 (5k) c48 (5k) c50 (5k)

Note. Cluster number shown as "c4 (1k)," meaning Cluster No. 4 of the first 1000-word band.

knowledge, as demonstrated by Laufer and Goldstein (2004) and Schmitt (2014). Naoko began cluster c26 with the meaning of "agree," the first meaning listed in the cluster's left column. Another example is cluster c2, where Naoko says "Body part that sees. It's only 'eye'. Parent who is a man. Parents are only fathers or mothers, so… part of the day with no sun… it means cloudy or just night… It's night" (Think-aloud, line 2). Each of these instances started with the term on the left read first and then matched to the correct word form listed at the top.

In total, 84% of the even-numbered clusters (21 of 25) saw at least partial use of this strategy of reversed matching (Table 2). Switching that was not verbalized or noticed could have occurred for the four other clusters as well, but the full extent of switching is not observable. It is also possible to repeatedly switch. Bidirectional matching is at odds with the measuring of written receptive VK employable for reading because definitions are not provided first with words presented second in the act of receiving written texts. Although the aspects of vocabulary are interrelated, the accuracy of estimating pertinent VK is reduced when vocabulary measures and language performance measures are not ecologically aligned (McLean et al., 2020; Schmitt, 2014).

Another affordance that Naoko acted upon was eliminating options in order to solve for items. Table 3 details the observed behavior of option elimination and guesswork on clusters listed by the number of correct answers. The observed data suggested that she used elimination to varying degrees by cluster. Also, since each cluster was unique, Naoko's partial knowledge specific to each cluster could have inflated her number of correct answers in idiosyncratic ways. From the think-aloud and the interview, it appeared that 72% of the even-numbered clusters (18 of 25) were answered using elimination. The incorrect answers for clusters c28 and c50 appeared to be total guesses. Clusters c2, c4, c6, and c8 appeared to be correctly answered without eliminating options. The clusters with no evidence of option elimination appeared to be clusters of words and definitions that were very easy or very difficult to understand for Naoko. Nonetheless, unobserved or unverbalized option elimination could have also been present, which would increase the weight of evidence.

As Wilson (1994) outlined, not all conscious thought is expressed in a think-aloud protocol. Additionally, not all recollections are expressed during an interview. Therefore, the actions of switching and elimination could have occurred to

an even greater extent, but even so, such unobserved action does not harm the evidence that was collected. Naoko was observed to switch matching direction between a mode of processing meaning from form and form from meaning. She was also observed to eliminate options to solve for clusters. Because Naoko did not seem to eliminate options to solve for the very easiest or the very hardest clusters, it may be that she was prompted to act at a threshold of difficulty but had to act within the limits of her knowledge of the given lexis. In taking the theoretical position of Gibson (1979) and Norman (1988), the bounds of this evidence invite the possibility that, when a test taker takes up the actions of switching matching direction and answering option elimination, it is the human response to this testing environment.

5 Discussion

5.1 Building a Taxonomy for Testing Receptive Vocabulary Knowledge

Naoko's answers for target words on the UVLT suggested a hierarchy of action. It seemed warranted to organize these actions within a parsimonious structure. Actions repeated here from the code book were differentiated as follows:

Strategies afforded by test format

a. *Guess an answer for unknown reason*

b. *Guess an answer from word part*

c. *Guess an answer from loanword knowledge*

d. *Select an answer by eliminating other options*

e. *Switching between matching word target to answer option and vice versa*

A taxonomic structure emerged from the categorization of observed actions and VK. Patterns began to fit the affordances inferred to exist for the other commonly used written receptive vocabulary testing formats summarized above.

Saldaña (2016) detailed how a taxonomy can represent the actions of people in a setting. Spradley (1980) illustrated different forms that taxonomies take in social settings, such as tree diagrams, but he also described the actions of the lone person or even the functionality of an object by using taxonomic diagrams of various configurations (see pp. 114 & 120). In this tradition, the observed patterns were delineated as an expanding order of test-taking actions made possible by format. A taxonomy emerged that structured the affordances provided by the clustered form-meaning matching format. The meaning-recognition format was reasoned to not afford as many test-taking actions as the item-clustered format. The meaning-recall format was included but was reasoned to be most restrictive of test-taking actions.

In the *Taxonomy of Test-taking Actions Afforded by Receptive Vocabulary Test Format* (see Figure 3), affordances provided by item format are illustrated by three shaded boxes, each expanding outward from smallest to largest in size.

Table 3. Option elimination and correctly answered items per cluster during think-aloud protocol

	Correct target items per 3-word cluster			
Observed behavior	3 (all) correct	2 correct	1 correct	None correct
Complete guesswork No elimination observed Stated to not know any words	c26 (3k)			c28 (3k) c50 (5k)
(Eliminating at least one option) Guess between options (>2)	c12 (2k) c14 (2k) c32 (4k) c38 (4k) c40 (4k) c44 (5k)	c22 (3k) c46 (5k)	c24 (3k) c34 (4k) c36 (4k) c48 (5k)	c30 (3k)
(Eliminating all but one option) Select the remaining option	c10 (1k) c18 (2k) c20 (2k) c42 (5k)	c16 (2k)		
Directly select option No elimination observed	c2 (1k) c4 (1k) c6 (1k) c8 (1k)			

Note. Cluster number shown as "c4 (1k)," meaning Cluster No. 4 of the first 1000-word band.

The various test strategies detailed in this study are all bounded within the largest box representing the clustered form-meaning matching format of the UVLT. This item-clustered format introduces a decision point to the test taker: the ability to switch between matching target form to target meaning and target meaning to target form. Naoko took up this affordance of switching matching direction, as shown in Table 3.

When matching target form to target meaning, the path in the taxonomy enters the middle box representing the meaning-recognition test format. This format allows the test taker to decide whether to eliminate options among a list of definitions using VK, even if partial or incomplete. The action of eliminating options is also afforded to test takers who switch their matching direction, but such action remains in the outer box because it would not belong within the middle box representing meaning-recognition testing (e.g., Schmitt et al., 2020).

The decision points are connected by arrows on the taxonomy. Switching could be repeated between the matching directions afforded by item-clustered test format. Switching could happen and return back unnoticed if not verbalized. For the meaning-recognition test format, the arrows within the box illustrate the logical result that once an option is eliminated, this action is not reversible: one could second-guess themselves, but a state of "yes" remains for that cluster. If the test taker does not act on the affordance of eliminating options, a test taker directly selects an option.

However, the presence of option choice keeps test-taking action outside the innermost box (meaning-recall test format). As a result, the taxonomy has no arrows entering the innermost box, which excludes the possibility of test takers to ignore the affordances introduced by the meaning-recognition test format and the

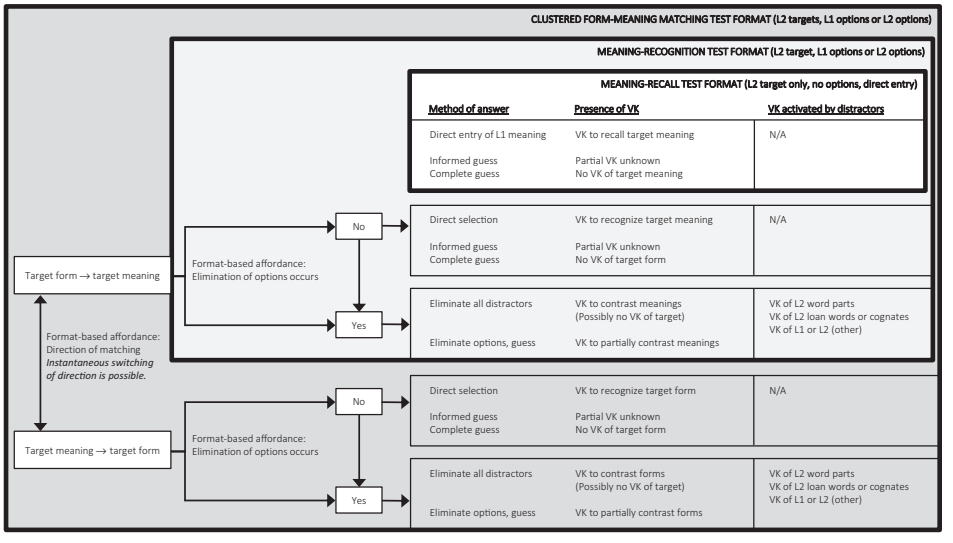


Figure 3. Taxonomy of test-taking actions afforded by receptive vocabulary test format.
Note. VK = Vocabulary knowledge.

clustered form-meaning matching format. The meaning-recall test format only allows the test taker to enter L1 meaning for an L2 word form directly into a blank answer space. In other words, affordances provided by the formats of the outer two boxes preclude the ability to demonstrate written receptive VK at the level of meaning recall.

5.2 Heuristic for Evaluating Item Format

Emerging from this study's findings, the *Taxonomy of Test-taking Actions Afforded by Receptive Vocabulary Test Format* in Figure 3 illustrates how researchers can narrow their L2 vocabulary measurement to knowledge that they theorize to be relevant to their studies. A review of the test formats entered into the taxonomy is warranted. Functioning within the middle box of the taxonomy is the meaning-recognition item format of multiple-choice testing. Although its purpose is not the same as the VLT, the VST by Beglar (2010), detailed in the literature review, is an example of a meaning recognition test (Figure 4). This format does not afford matching or bidirectional matching. The innermost box in Figure 3 represents meaning-recall testing (Figure 5), which does not afford option selection, nor option elimination, nor option matching (e.g., vocableveltest.org [McLean & Raine, 2019]).

Naoko's test taking exhibited behaviors centered on eliminating options and the switching of form-meaning matching direction. Neither of these test-taking strategies aid in the processing of L2 input when actually reading. Given the affordance, test takers utilize partial knowledge to solve for items. The UVLT and its predecessors are often used as a measure in experimental research designs. As discovered in this study, test results can be realized in ways unknown and unintended to testing researchers' purposes. Nevertheless, an accurate measure of

1. miniature: It is a miniature.
   a. a very small thing of its kind
   b. an instrument for looking at very small objects
   c. a very small living creature
   d. a small line to join letters in handwriting

Figure 4. An example item from the meaning-recognition format of the VST.

1. time: They have a lot of time
< direct input of L1 meaning by test taker>

Figure 5. An example item from a meaning-recall test format.

VK is important for experimental research and pedagogic aims with L2 learners. For example, VLT has been used to match L2 learners with lexically appropriate reading materials, and in the case of fluency building, 98–100% of the words should be known to ensure that lexical difficulty does not impede the processing of L2 input (Hu & Nation, 2000). Vocabulary test scores allow for estimates of lexical coverage in this case, which include an inherent margin of error, but a test maker should develop or select a test to reduce this error as much as possible.

Precision in estimating VK is ideal to avoid selecting L2 materials that are too easy or too difficult for a research or pedagogic purpose. Within a framework such as the four strands (Nation, 2007), there are additional pedagogic aims for known vocabulary in texts. Erroneous estimations of lexical coverage for study materials could find texts intended for fluency building to be too difficult and actually be suited for form-focused instruction (i.e., lexical coverage at 95%). The difference between 98% and 95% of lexical coverage equates to a difference of one or two extra words per 50 words not being known by an L2 learner. This narrow margin for error would be sensitive to an inappropriate testing format. The taxonomy outlines the influence of affordances that item format introduces. On the basis of vocabulary studies concerning item format that have been cited in this study and the observation data of this study, the proposed taxonomy can serve as a reference for researchers to further explore relevant issues of vocabulary assessment and development.

6 Conclusion

This study is rooted in data about the functioning of the UVLT gathered from a think-aloud protocol and retrospective interviewing. Qualitative and test score data were coded and analyzed. Patterns of affordances emerged, which led to a taxonomy of test-taking actions argued to be made possible to test takers by test item format. The expanding boxes of the taxonomy represent additional affordances that increase test-takers' abilities to make use of degrees of VK and guesswork. In fact, the clustered form-meaning matching format (outermost box)

afforded instances of correctly answering for target items within clusters despite having no apparent knowledge of the target words. At three levels in the taxonomy (item-clustered format, meaning-recognition format, and meaning-recall format), VK inferred to be usable at each level is detailed.

The presence of affordances and use of partial VK narrows at each level in the taxonomy in a graphic form. It illustrates a hierarchy of testing conditions used in empirical studies about the effects of item format on measuring VK (Kremmel & Schmitt, 2016; Laufer & Goldstein, 2004; McLean et al., 2020). Furthermore, the hierarchy of test scores in these studies amounts to a convergence of evidence that format matters for written receptive vocabulary testing. When a test taker performs the actions of switching, matching, and eliminating, these are human responses to the testing environment. In sum, the taxonomy proposed in this exploratory study provides a structure from which item format can be viewed and evaluated given pedagogic or research aims of interest.

Acknowledgments

I would like to thank Dr. Tamara Swenson and other colleagues, as well as the two anonymous reviewers, for their constructive comments and suggestions, all of which helped improve the manuscript.

References

Atkinson, D., Churchill, E., Nishino, T., & Okada, H. (2018). Language learning great and small: Environmental support structures and learning opportunities in a sociocognitive approach to second language acquisition/teaching. *The Modern Language Journal, 102*(3), 471–493. https://doi.org/10.1111/modl.12496

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing, 27*(1), 101–118. https://doi.org/10.1177/0265532209340194

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and university word level vocabulary tests. *Language Testing, 16*(2), 131–162. https://doi.org/10.1177/026553229901600202

Bond, T. G., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge.

Churchill, E., Okada, H., Nishino, T., & Atkinson, D. (2010). Symbiotic gesture and the sociocognitive visibility of grammar in second language acquisition. *The Modern Language Journal, 94*(2), 234–253. https://doi.org/10.1111/j.1540-4781.2010.01019.x

Gibson, J. J. (1979). *The ecological approach to visual perception*. Erlbaum.

Ha, H. T. (2022). Test format and local dependence of items revisited: A case of two vocabulary levels tests. *Frontiers in Psychology, 12*, 805450. https://doi.org/10.3389/fpsyg.2021.805450

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403–430.

Johnstone, C. J., Bottsford-Miller, N. A., & Thompson, S. J. (2006). *Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners* (Technical Report 44). University of Minnesota, National Center on Educational Outcomes. https://nceo.info/Resources/publications/OnlinePubs/Tech44

Kamimoto, T. (2014). Local item dependence on the vocabulary levels test revisited. *Vocabulary Learning and Instruction, 3*(2), 56–68. https://doi.org/10.7820/vli.v03.2.kamimoto

Kremmel, B., & Schmitt, N. (2016). Interpreting vocabulary test scores: What do various item formats tell us about learners' ability to employ words? *Language Assessment Quarterly, 13*(4), 377–392. https://doi.org/10.1080/15434303.2016.1237516

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning, 54*(3), 399–436. https://doi.org/10.1111/j.0023-8333.2004.00260.x

McLean, S., & Raine, P. (2019). *VocabLevelTest.Org* [Web application]. https://www.vocableveltest.org

McLean, S., Stewart, J., & Batty, A. O. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing*, 026553221989838. https://doi.org/10.1177/0265532219898380

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines, 5*(1), 12–25.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Heinle & Heinle.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*(1), 59–82.

Nation, I. S. P. (2007). The Four Strands. *Innovation in Language Learning and Teaching, 1*(1), 2–13. https://doi.org/10.2167/illt039.0

Nation, I. S. P. (2012). *The BNC/COCA word family lists*. http://www.victoria.ac.nz/lals/about/staff/paulnation

Nation, I. S. P. (2020). The different aspects of vocabulary knowledge. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (1st ed., pp. 15–29). Routledge.

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary* (1st ed.). Heinle, Cengage Learning.

Norman, D. A. (1988). *The psychology of everyday things*. Basic Books.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danmarks Paedogogiske Institut.

Read, J. (2020). Key issues in measuring vocabulary knowledge. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (1st ed., pp. 545–560). Routledge. https://www.taylorfrancis.com/books/9781000005561/chapters/10.4324/9780429291586-34

Saldaña, J. (2016). *The coding manual for qualitative researchers* (3rd ed.). Sage.

Schmitt, N. (1999). The relationship between TOEFL vocabulary items and meaning, association, collocation, and word-class knowledge. *Language Testing, 16*(2), 189–216. https://doi.org/10.1177/026553229901600204

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows: Size and depth of vocabulary knowledge. *Language Learning, 64*(4), 913–951. https://doi.org/10.1111/lang.12077

Schmitt, N., Nation, I. S. P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching, 53*(1), 109–120. https://doi.org/10.1017/S0261444819000326

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*(1), 55–88. https://doi.org/10.1177/026553220101800103

Smith, E. V., Jr. (2001). Evidence for the reliability of measures and the validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement, 2*(3), 281–311. http://jampress.org/

Spradley, J. P. (1980). *Participant observation*. Waveland Press.

Van Lier, L. (2004). *The ecology and semiotics of language learning: A sociocultural perspective*. Kluwer Academic.

Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL—International Journal of Applied Linguistics, 168*(1), 33–69. https://doi.org/10.1075/itl.168.1.02web

Wilson, T. D. (1994). The proper protocol: Validity and completeness of verbal reports. *Psychological Science, 5*(5), 249–252. https://doi.org/10.1111/j.1467-9280.1994.tb00621.x