An Item Response Theory analysis of the Verb Subordinates Test

Aleka Akoyunoglou Blackwell & Jennifer Flipse

Middle Tennessee State University
Aleka.Blackwell@mtsu.edu, *Jennifer.Flipse@mtsu.edu*

Abstract
This study presents a psychometric evaluation of the Verb Subordinates Test (VST). The VST assesses lexical competence based on knowledge of troponyms in the verb lexicon. Items are true/false statements with the structure *To verb*$_{hyponym}$ *is a way to verb*$_{hypernym}$. Using Item Response Theory (IRT) analysis, this study examined the difficulty and discriminatory value of different items and difficulty levels of the VST. Statistical analyses showed that the VST is a promising vocabulary assessment measure with high internal consistency and good convergent validity, and that individual VST items, given their frequency range, are differentially informative across the vocabulary trait continuum.

Keywords: vocabulary assessment, Verb Subordinates Test, Item Response Theory, hypernymy, hyponymy, superordinates, subordinates, depth of lexical knowledge

1 Introduction

In a hierarchical model of verb semantics, pairs of verbs such as *nibble* and *eat*, *stutter* and *talk*, and *traipse* and *walk* are in a subordinate/superordinate relation with each other. This semantic relationship between verbs has been aptly termed *troponymy* by Fellbaum & Miller (1989). The term is derived from the Greek word *τρόπος* ('way, manner') to reflect the fact that it is specifically a manner relation, i.e., *To V*$_1$ *is to V*$_2$ *in some particular manner* (Fellbaum 1998). Within the hierarchical network theory of the mental lexicon reflected in WordNet (Miller 1990), the troponymy relation has given rise to a unique type of taxonomic hierarchy in the representation of the verb lexicon. The hierarchy is characterized by a "shallow, bushy structure" with typically only four hierarchical levels and "what might be called a bulge, that is to say, a level with far more verbs than the other levels in the same hierarchy" (Fellbaum 1998: 80). For example, the hierarchy for the verb *talk* (in the sense 'express in speech') is headed by the verb *communicate* (in the sense 'transmit thoughts or feelings'), includes a small set of sister verbs (e.g., *inform, talk, write, share*), and bulges at the subsequent level (the level of troponyms of *talk,* such as, *shout, whisper, babble, rant, mumble, chatter, slur, bark, hiss, sing, peep, whiff, blubber, drone, rasp, yack, murmur, snivel, cackle, blurt out, verbalize, lip off, speak up, troll*).

Aside from being an organizing principle of the verb lexicon in WordNet, the troponymy relationship is also reflected in the organization of verbs in the mental lexicon. In tasks involving semantic processing, troponymy has a unique status among semantic relations, including opposition and synonymy. Troponymy is the most frequent semantic relation elicited in word association tasks involving verb responses to a verb stimulus (Fellbaum & Chaffin 1990), and it is the dominant relation guiding behavior in analogy and sorting tasks (Chaffin et al. 1994 as cited in Fellbaum 1998) as well as in elicitation tasks of semantic commonalities between verb pairs (Pavličić & Markman 1997).

A distinction between troponyms and their superordinates is also reflected in the order of acquisition of verbs in both first and second language. In the early stages of first language acquisition, children rely initially on General All-Purpose verbs (GAP verbs), such as *do, put, get, come, go, make*, whereas troponyms are acquired more slowly and gradually (Kambanaros & Grohmann 2015; Rice & Bode 1993; Thordardottir & Ellis Weismer 2001). Second language research on the acquisition of the verb lexicon has replicated this finding, both in immersion programs (Harley 1992) and in traditional L2 contexts (Crossley 2013; Crossley et al. 2009).

These different lines of research converge in suggesting that the troponymy semantic relation in the verb lexicon has psychological validity. Furthermore, the developmental evidence suggests that the hierarchical organization of the verb lexicon is also reflected in the verb acquisition trajectory, with GAP verbs, i.e., the level which parallels the basic level in noun hierarchies (Rosch et al. 1976), being acquired prior to the bulging level of troponyms. In line with this model of the acquisition and organization of the verb lexicon, Blackwell (2012) developed a vocabulary assessment, the Verb Subordinates Test (VST), which relies on the hypernym/troponym semantic relation. In this study, we investigated the psychometric properties of the test as well as the individual test items. The goals were to determine (i) whether vocabulary test items relying on the presence or absence of troponymy in verb pairs are effective at discriminating individuals across different levels of lexical competence, and (ii) which difficulty levels of the VST are most informative at different levels of lexical competence.

Traditionally, the development and validation of vocabulary tests has been guided by classical test theory (CTT). By contrast, item response theory (IRT), an alternate measurement framework, is specifically designed to evaluate individual test items in terms of their difficulty as well as their ability to discriminate between test-takers of different proficiency levels (e.g., Hoffman et al. 2012), and it has been used to validate many standardized measures, particularly in the field of computer adaptive testing (e.g., Kingsbury & Houser 1993). We, therefore, employed an IRT analysis in this study.

The paper proceeds as follows: section two introduces the Verb Subordinates Test and discusses methodological considerations guiding its development; section three presents an overview of Item Response Theory, including relevant IRT models and assumptions; section four describes the research methodology. The remaining sections present the results of the IRT analysis, a discussion of three IRT models, and conclusions on the effectiveness of the VST as a vocabulary assessment measure.

2 The Verb Subordinates Test

The Verb Subordinates Test (VST) consists of 40 test items. The items represent five levels of difficulty with eight items per difficulty level. The items are all true/false statements seven words in length. Each item has the structure [To verb$_x$ is a way to verb$_y$] where verb$_x$ is a troponym of verb$_y$. The target verbs in the VST are by definition a troponym, i.e., the definition of the test verb in WordNet includes its hypernym, e.g., the definition of *trundle*, the selected troponym of *move*, includes the verb *move* (“to move heavily”). In addition, each hypernym appears in only one level on the test, once in a true statement and once in a false statement. For example, the hypernym *jump* appears only in Level 1 in the items *To bounce is a way to jump* (true) and *To sip is a way to jump* (false); the hypernym *talk* appears only in Level 2 in the items *To rasp is a way to talk* (true) and *To slurp is a way to talk* (false). Lastly, the selected troponyms have at most two senses in WordNet. For example, the verb *prate* has one sense ("speak about unimportant matters rapidly and incessantly"), and the verb *roast* has two senses ("to cook in dry heat, usually in the oven" and "to subject to laughter or ridicule").

The difficulty levels on the VST are based on target verb frequency in the Corpus of Contemporary American English (COCA) (Davies 2011). The target verbs in level 0, the simplest level, are all morphologically derived from their hypernyms (e.g., overeat/eat, outgrow/grow). The target verbs in levels 1-4 are drawn from decreasing frequency ranges in COCA (see Table 1). In all 40 items, the hypernyms are high frequency, familiar, basic level verbs (GAP verbs). The VST appears in its entirety in the appendix.

ST Difficulty Levels	**Troponyms**
Level 0 (target verb is morphologically derived from its hypernym)	*overhear, remake, misfire, outgrow, sleepwalk, handwrite, spoonfeed, outrun*
Level 1 (target verb within top 7.5K lemmas in the 60K list of lemmas in *COCA*)	*devour, jog, roast, chant, bounce, sip, chop, hop*
Level 2 (target verb between 18K and 23K lemmas in the 60K list of lemmas in *COCA*)	*trundle, core, beseech, wend, lope, guzzle, rasp, slurp*
Level 3 (target verb between 30K and 45K lemmas in the 60K list of lemmas in *COCA*)	*burgeon, jounce, hanker, flub, quaff, dodder, snivel, swill*
Level 4 (target verb less frequent than the top 60K lemmas in the 60K list of lemmas in *COCA*)	*reave, prate, gawp, saltate, lollop, piffle, pronk, scarper*

Table 1. VST target verbs by level

3 Overview of Item Response Theory

IRT is a statistical procedure that was developed to model the relationship between the construct being measured by a test and the individual items on the test. For each item, IRT provides an item characteristic curve (ICC) which graphs the probability that a test-taker will answer an item correctly given their ability level. In IRT, ability (or *latent trait* in IRT terminology) is represented by the variable theta (θ). In an ICC plot (see Fig. 1), θ is represented along the *x*-axis and usually ranges between -3 and +3, with 0 representing average ability level. The probability of a correct response on an item is graphed on the *y*-axis and is scaled from 0.0 to 1.0.

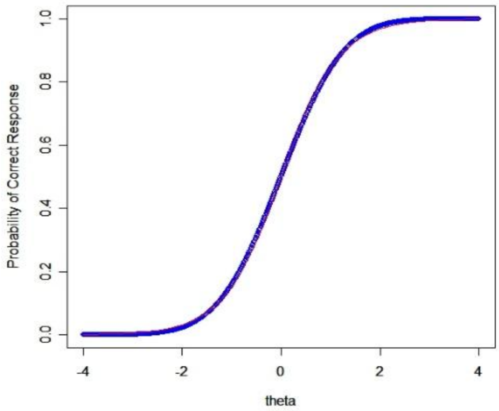

Figure 1. Item Characteristic Curve

The steeper the ICC curve, the better the represented item discriminates between test-takers with contiguous trait levels. An item's *discrimination parameter* is denoted by *a*, and it is defined as the slope of the ICC at an item's difficulty (or *location parameter* in IRT terminology). An item's difficulty is denoted by *b* and represents the ability level required for a test-taker to have a .50 probability of answering the item correctly.

A variety of specific IRT models have been developed based on (i) the number of item characteristics (or *parameters* in IRT terminology) included in the model and (ii) the type of test item (dichotomous measures vs. polytomous measures). We focus here on the three models appropriate for dichotomous measures, as is the case with VST items. The simplest such model is the *Rasch Model*, also known as the *one-parameter logistic model* (1PL) (Rasch 1960). This model estimates the difficulty of each item assuming a constant discrimination parameter across all items. By comparison, the more complex *two-parameter logistic model* (2PL) estimates both the difficulty of each item and its discrimination parameter. Lastly, with dichotomous items where guessing can be a significant factor in performance, IRT provides a -*three-parameter logistic model* (3PL) which takes into account item difficulty, item discrimination, and a guessing parameter.

All IRT models rely on four assumptions. The first is unidimensionality, i.e., the assumption that all items measure the same, single latent trait. The second is local independence, i.e., each item on a test is statistically independent of responses to all other items on the measure. The third is monotonicity, i.e., the expectation that the probability of endorsing an item will continuously increase as an individual's trail level increases. The fourth is item invariance, i.e., the assumption that estimated item parameters are constant across different populations. If this last assumption is not supported by the data, the IRT analysis provides information on how different items behave with different subgroups of the population after controlling for ability.

Lastly, the sample size requirements for IRT parameter estimations vary based on the choice of model and type of items, with ranges from 100 to 500 participants.

4 Method

4.1 Participants

Three hundred and five participants were recruited from the undergraduate psychology pool of a large university in the United States (Age, *Min* = 18, *Max* = 41, *Mode* = 19, *Mdn* = 19, *M* = 20.63, *SD* = 3.44). All were native speakers of English, had normal or corrected-to-normal vision and normal hearing, and participated for course credit. The sample was deemed representative of the general population because participants were largely freshmen enrolled in a nonselective university, and their lexical competence was expected to vary adequately for the psychometric properties of the VST to be evaluated. The number of participants was deemed appropriate in light of sample size recommendations for IRT analysis (Stone & Yumoto 2004). The study was approved by the MTSU Institutional Review Board, and all participants provided written informed consent.

4.2 Materials

Materials included the VST and the Peabody Picture Vocabulary Test (PPVT-III) (Dunn 1997) which served as the standardized measure of vocabulary size and a measure of concurrent validity for the VST.

4.3 Procedure

The VST was administered on a computer. Items were presented electronically using the E-Prime 2.0 software (Psychology Software Tools, Pittsburgh, PA). Items appeared one at a time and centered on the computer screen. Participants were instructed to press the *A* key on the keyboard if they thought the statement on the screen was true or the *L* key on the keyboard if they thought the statement on the screen was false. Participants were instructed to respond to each item as quickly and accurately as possible. Each item remained on the screen until the participant pressed the chosen key. After each item response, participants reported how confident they felt about the accuracy of their answer on a scale of 1 to 5. The 40 test items appeared in a randomized order. Participants completed a set of three practice trials to become familiar with the procedure. The entire test lasted approximately 15 minutes. The PPVT-III was administered individually in a quiet private setting by a trained experimenter. Participants completed this task in approximately 20 minutes. The order of the two tests was counterbalanced between participants.

4.4 Statistical analyses

Descriptive statistics examined the range of scores on VST as well as on each of the difficulty levels of the test. Cronbach's alpha measured internal consistency. Pearson product-moment correlations between the VST and the PPVT-III evaluated convergent validity. IRT analyses, conducted with *XCalibre* 4.1 (Guyer & Thompson 2012), explored model and item fit in the 1PL, 2PL, and 3PL IRT models. IRT analyses were also conducted excluding misfitting items. Finally, IRT analyses were conducted on different subsets of test items representing different difficulty levels in order to determine the least number of test items that maximized the information provided by the VST across the widest range of lexical competence.

5 Results

5.1 Vocabulary profile of the sample

The PPVT-III served as standardized measure of vocabulary ability. As expected for a sample drawn from a college student population, the mean of the age-normalized PPVT scores in the sample was above the mean of the normative sample, and the standard deviation was smaller than that of the normative sample (sample standardized *M* = 106.34, *SD* = 10.29). However, the range of the standardized percentile rank of the participants in our sample was large (Percentile Rank Range: 7 – 99, *Median* = 66). The sample was, therefore, deemed appropriately representative and adequately diverse in its range of lexical competence to be used in a psychometric evaluation of the VST.

5.2 Descriptive statistics, internal consistency, and convergent validity

Sample means, standard deviations, and score ranges on the PPVT-III and VST appear in Table 2. The VST has good internal consistency (Cronbach's alpha = .871), and it remains a reliable assessment measure across the combinations of difficulty levels examined (α_{range} = .871 – .905) (see Table 2 for combinations of difficulty levels analyzed). Convergent validity was established based on a correlational analysis between the VST and the PPVT-III which revealed a moderate, positive, significant correlation between the VST and both the PPVT-III raw score (*r* = .495, *p* < .01) and the PPVT-III standard score (*r* = .486, *p* < .01).

Measure	**Details**	***Mean***	***SD***	***Range***
PPVT-III	Standard score	106.34	10.29	78 – 138
Raw score range= 0-204	Raw score	177.73	10.43	133 – 200
VST	Level 0	7.59	.69	5 – 8
Score range= 0-8	Level 1	7.79	.54	5 – 8
	Level 2	6.33	1.21	3 – 8
	Level 3	5.01	1.34	1 – 8
	Level 4	4.76	1.24	1 – 8
VST (level subsets)				
Levels 0-4	40 items	31.40	3.22	11 – 39
Levels 0-3	32 items	26.66	2.70	11 – 32
Levels 1-4	32 items	23.81	3.02	16 – 31
Levels 2-4	24 items	16.05	2.73	9 – 23

Table 2. PPVT-III and VST means, standard deviations, and ranges

5.3 IRT models

Comparisons of IRT models of the full 40-item VST revealed that the 3PL model (which takes into account item difficulty, item discrimination, and guessing parameters) with misfitting items removed provides the overall best fit indices (*-2LL* = 20389, χ^2 = 1662.80, *p* = .113). Examination of the test information function curve of the 3PL model indicated that the 40-item VST is most effective at assessing vocabulary knowledge for participants with vocabulary knowledge one standard deviation below the mean (θ = -.006±.95) and provides maximum information at θ = -1.3 (see Figure 2).

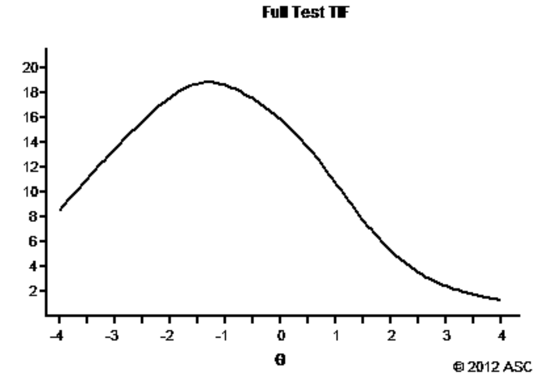

Figure 2. 40-item 3PL model test information function[1]

IRT analyses of subsets of difficulty levels revealed that the 3PL model of the set of 24 items in difficulty levels 2 – 4 provided the maximum discrimination information across the largest range of theta values (θ_{range} = -2.0 – +1.5) (see Figure 3). This model also had good reliability (α = .898) and appropriately centered theta values (θ = -.006±.96). Model fit statistics for this model were *-2LL* = 25887 and χ^2 = 1988.37, *p* = <.001.

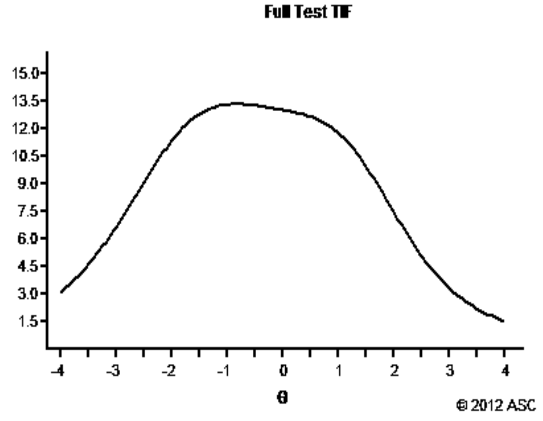

Figure 3. 24-item 3PL model test information function

[1] The Test Information Function (TIF) represents the relative precision of the test across different levels of the trait continuum, and the height of the TIF is proportional to the standard error of measurement.

The 3PL model of the set of 32 items in difficulty levels 1 – 4 had good reliability (α = .904) with appropriate centered theta estimates (θ = -.006±.95). This model provided maximum information theta levels that were between the mean theta and one standard deviation below the mean (θ = -.75) (see Figure 4).

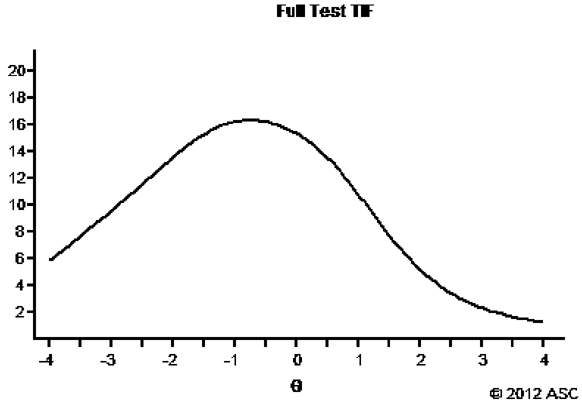

Figure 4. 32-item 3PL model test information function

5.4 Item results

IRT analysis allows us to identify more discriminating items, that is, items which provide greater information about a respondent, from less discriminating items which are not as informative. The best and worst items in this regard for each level of the test appear in Table 3. As mentioned earlier, for each item, IRT also provides an item characteristic curve (ICC) which graphs the probability that a test-taker will answer an item correctly given their ability level. The steeper the ICC curve, the better the represented item discriminates between test-takers with contiguous trait levels. We present here an example of the ICC of a "good" test item (the item *To core is a way to move*) in Figure 5 and an example of the ICC of a "poor" test item (the item *To burgeon is a way to grow*) in Figure 6.

	Best Item	χ^2	*p*	**Worst Item**	χ^2	*p*
L0	To remake is a way to feed	4.71	.967	To outrun is a way to grow	16.45	.172
L1	To chop is a way to cut	5.60	.935	To hop is a way to cut	17.11	.145
L2	To beseech is a way to ask	4.09	.982	To lope is a way to run	32.91	<.001
L3	To rasp is a way to talk	7.54	.820	To core is a way to move	17.75	.124
L4	To pronk is a way to jump	7.41	.829	To prate is a way to take	21.31	.046

Table 3. Best and worst items at each difficulty level

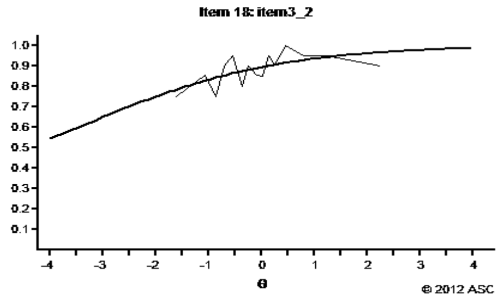

Figure 5. Characteristic Good Item Fit

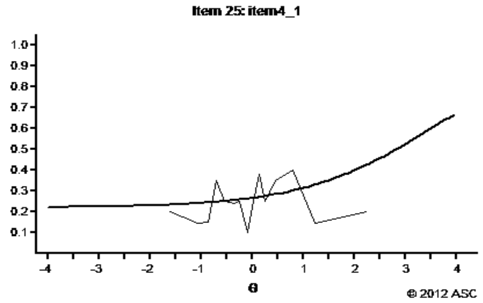

Figure 6. Characteristic Poor Item Fit

6 Discussion

The goal of this study was two-fold. The first goal was to evaluate the VST, including the semantic relation of troponymy in the verb lexicon, as a measure of vocabulary knowledge. The second goal was to determine whether a subset of difficulty levels relying on verb frequency in COCA is more or less informative as a measure of vocabulary knowledge for test-takers at different levels of lexical competence.

Our analyses showed that the VST is a promising vocabulary assessment measure with high internal consistency and good convergent validity. The significant positive correlation between the VST and the PPVT-III offers convincing evidence that assessing lexical competence by evaluating test-takers' knowledge of verb pairs which are related by means of troponymy is a promising endeavor. Moreover, the moderate rather than strong correlation between the PPVT and the VST suggests that the VST measures additional aspects of vocabulary knowledge beyond those measured by the PPVT. In other words, this type of test may, in fact, "make a substantial contribution to assessing the state of a learner's vocabulary knowledge beyond what is measured by a well-designed test of vocabulary size" (Read 2004: 224).

In terms of the psychometric properties of the VST, IRT analyses revealed that items on the VST are differentially informative and differentially successful at discriminating different levels of the latent trait. In addition, the results suggest that true items maybe more informative than false items (see Table 3 where four out of the five best items are true statements whereas four of the five worse items are false statements), and this finding certainly warrants further investigation. Finally, IRT analyses revealed that different combinations of difficulty levels are most informative with different populations. Specifically, with a population of college students (i.e., the population from which our sample was drawn), only the items in difficulty levels 2-4 are necessary when using the VST to assess their vocabulary knowledge. If, however, one were to use the VST to assess a population whose vocabulary knowledge is a standard deviation below that of our sample, then levels 1-4 would be most informative for that population. For a population whose vocabulary knowledge is expected to be lower than that, then levels 0-4 would need to be administered to adequately assess these individuals. As we see it, the shorter the test without sacrificing the validity, reliability, or information function of the test, the better in terms of conserving resources in assessment.

One major advantage of the VST is that it is possible to create new items using WordNet and COCA. The test itself is also easy to administer, and it is easy to score. This type of vocabulary test can, therefore, be easily tailored to specific contexts, both in first and second language settings, with both adults and child learners of English. Our study focused on exploring the psychometric properties of the VST as a vocabulary measure for young adult native speakers of English. It seems likely that the VST can also be used in ESL and ELL contexts; however, its psychometric properties with such populations must be similarly established.

Lastly, we believe this study makes at least two important contributions. First, it validates a new approach for assessing vocabulary knowledge, not only by incorporating the troponymy relation between verb pairs in vocabulary testing, but also by focusing specifically on the verb class as a means to assess lexical aptitude. Second, it adds new evidence from L1 adult vocabulary testing to the line research on the status of hypernymic/hyponymic relationships both in the developing L1 child lexicon (Mervis & Crisafi 1982; Murphy 2004) and in the growth of the L2 lexicon (Crossley 2013; Crossley et al. 2009; Haastrup & Henriksen 2000; Sharifian 2002).

7 Conclusion

This paper presented a psychometric evaluation of the Verb Subordinates Test. The VST represents a new methodology for assessing lexical competence, and our evidence suggests that this approach to assessing lexical aptitude is promising. The

IRT analyses revealed that individual vocabulary items on this measure, given their frequency range, are differentially informative across the vocabulary trait continuum. With that in mind, future uses of this measure can tailor items based on whether the goal is to discriminate people on the high end of the trait continuum vs. the lower range of the continuum. Furthermore, the VST offers the advantage of easy administration, and new items can be readily developed using WordNet and COCA. Lastly, verb subordinate tests of this type can be developed as vocabulary assessment measures in any language, provided the availability of corpus-based frequency data of verb use by its speakers.

References

Blackwell, A.A. (2012). *The Verb Subordinates Vocabulary Test*. Unpublished manuscript, Middle Tennessee State University.

Crossley, S. (2013). Assessing automatic processing of hypernymic relations in first language and advanced second language learners. *The Mental Lexicon* 8(1): 96-116.

Crossley, S., T. Salsbury & D. McNamara (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning* 59(2): 307-334.

Davies, M. (2011). The corpus of contemporary American English as the first reliable monitor corpus of English. *Literacy and Linguistic Computing* 25: 447-465.

Dunn, L.M. (1997). *PPVT-III: Peabody picture vocabulary test*. Circle Pines, MN: American Guidance Service.

Fellbaum, C. (1998). A semantic network of English verbs. In C. Fellbaum (ed.), *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press, 69-104.

Fellbaum, C. & R. Chaffin (1990). Some principles of the organization of verbs in the mental lexicon. In M. Piattelli-Palmarini (ed.), *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates, 420-427.

Fellbaum, C. & G. Miller (1989). Folk psychology or semantic entailment? Comment on Rips and Conrad (1989). *Psychological Review* 97(4): 565-570.

Guyer, R. & N.A. Thompson (2012). *User's manual for Xcalibre item response theory calibration software, version 4.1.8*. St. Paul, MN: Assessment Systems Corporation.

Haastrup, K. & B. Henriksen (2000). Vocabulary acquisition: Acquiring depth of knowledge through network building. *International Journal of Applied Linguistics* 10(2): 221-240.

Harley, B. (1992). Patterns of second language development in French immersion. *French Language Studies* 2: 159-183.

Hoffman, L., J. Templin & M.L. Rice (2012). Linking outcomes from Peabody Picture Vocabulary Test forms using item response models. *Journal of Speech, Language, and Hearing Research* 55(3): 754-763.

Kambanaros, M. & K.K. Grohmann (2015). More general all-purpose verbs in children with specific language impairment? Evidence from Greek for not fully lexical verbs in language development. *Applied Psycholinguistics* 36: 1029-1057.

Kingsbury, G.G. & R.L. Houser (1993). Assessing the utility of Item Response Models: Computerized Adaptive Testing. *Educational Measurement: Issues and Practice* 12(1): 21-27.

Mervis, C.B. & M.A. Crisafi (1982). Order of acquisition of subordinate-, basic-, and superordinate-level categories. *Child Development* 53: 258-266.

Miller, G. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography* 3(4): 235-312.

Murphy, G.L. (2004). *The big book of concepts*. Cambridge, MA: MIT Press.

Pavličić, T. & A.B. Markman (1997). The structure of the verb lexicon: Evidence from structural alignment approach to similarity. In M.G. Shafto & P. Langley (eds.), *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates, 590-595.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* (Copenhagen, Danish Institute for Educational Research). Expanded edition (1980) with foreword and afterword by B.D. Wright, (1980). Chicago: The University of Chicago Press. Reprinted (1993) Chicago: MESA Press.

Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In P. Bogaards & B. Laufer (eds.), *Vocabulary in a second language: Selection, acquisition, and testing.* Amsterdam/Philadelphia: John Benjamins, 209-228.

Rice, M. & J.V. Bode (1993). GAPS in the verb lexicon of children with specific language impairment. *First Language* 13(37): 113-131.

Rosch, E., C.B. Mervis, W. Gray, D. Johnson & P. Boyes-Braem (1976). Basic objects in natural categories. *Cognitive Psychology* 8: 382-439.

Sharifian, F. (2002). Processing hyponymy in L1 and L2. *Journal of Psycholinguistic Research* 31(4): 421-436.

Stone, M. & F. Yumoto (2004). The effect of sample size for estimating Rasch/IRT parameters with dichotomous items. *Journal of Applied Measurement* 5: 48-61.

Thordardottir, E. & S. Ellis Weismer (2001). High-frequency verbs and verb diversity in the spontaneous speech of school-age children with specific language impairment. *International Journal of Language & Communication Disorders* 36: 221-244.

Appendix

Verb Subordinate Test Items

Level	*True Items*	*False Items*
0	To spoonfeed is a way to feed. To overhear is a way to hear. To handwrite is a way to write. To outgrow is a way to grow.	To remake is a way to feed. To sleepwalk is way to hear. To misfire is a way to write. To outrun is a way to grow.
1	To devour is a way to eat. To roast is a way to cook. To bounce is a way to jump. To chop is a way to cut.	To jog is a way to eat. To chant is a way to cook. To sip is a way to jump. To hop is a way to cut.
2	To trundle is a way to move. To beseech is a way to ask. To lope is a way to run. To rasp is a way to talk.	To core is a way to move. To wend is a way to ask. To guzzle is a way to run. To slurp is a way to talk.
3	To burgeon is a way to grow. To hanker is a way to want. To quaff is a way to drink. To snivel is a way to cry.	To jounce is a way to grow. To flub is a way to want. To dodder is a way to drink To swill is a way to cry.
4	To reave is a way to take. To gawp is a way to look. To lollop is a way to walk. To pronk is a way to jump.	To prate is a way to take. To saltate is a way to look. To piffle is a way to walk. To scarper is a way to jump.