A Rasch Look at Vocabulary Knowledge

Trevor HOLSTER, Miki TOKUNAGA, Simon WILKINS[i]

Abstract

Background. Disagreements arise over defining and testing vocabulary knowledge. "Trait" models define vocabulary knowledge independently of context, but "interactionalist" models see context as an important facet of knowledge. Other disagreements are over the relationship between "receptive" and "productive" translation direction, often assumed to measure different "dimensions" of knowledge, but with "receptive" knowledge preceding "productive" knowledge.

Aims. This research aimed to replicate Webb's (2008) investigation of the difficulty of translation direction, but used Rasch measurement to compare the difficulty of translation direction, of providing words in isolation and in a sentential context, and to investigate the dimensionality of "productive" and "receptive" translation tasks.

Sampling. Tests were administered to a convenience sampling of 260 learners with a mean TOEIC Bridge score of 111, a standard deviation of 16, and a range from 84 to 138, in the first semester of 2009.

.

Method. The 180 words tested by Webb (2008) were tested in English-to-Japanese and Japanese-to-English translation. Each word was tested without context and in a simple sentence. Responses in each direction were rated by two Japanese or two native speaker raters. Rasch measurement was used to equate the tests and compare persons, items, and raters on a common scale.

Results. "Productive" translation, from Japanese to English, was found to be considerably more difficult than "receptive", but unlike Webb (2008), no difference was found across proficiency levels. The two translation directions were not found to measure different dimensions of knowledge, but two facets of a unidimensional

construct. Importantly, presenting words in a sentence made a significant difference in translation difficulty compared with words presented in isolation.

Conclusions. Although knowledge of words presented in the L1 and L2 follow a single dimension of vocabulary knowledge, the manner and context of presentation of vocabulary are important facets of vocabulary knowledge. This casts doubt on trait views where vocabulary knowledge is defined in terms of decontextualized words. Rather, the results support interactionalist views of vocabulary knowledge, where emphasis is put on how words are used in context.

Background. 語彙力のテストや定義については見解の相違が存在する。「特性」論では語彙は文脈から独立したものとされるが、「相互作用」論では語彙の知識には文脈が重要な要因であるされる。他にも、受容的又は発表的の翻訳方向が語彙知識における異なる次元であり、受容的能力が発表的能力よりも先に習得されるという点においても議論がなされている。

Aims. 本研究は Webb(2008)の翻訳の方向による難易度の違いの検証を複製研究であり、Rasch 分析を行うことで翻訳方向や語彙が単独か文脈中かによる難易度への影響、そして受容的作業と発表的作業の次元的関係を検証する。

Sampling. テストは 260 名の英語学習者に対して 2009 年の前期に実施された。対象者の TOEIC Bridge スコアは 84 から 138 点（平均 111 点）、標準偏差は 16 であった。

Method. Webb (2008) で使用された 180 語を英日、日英の 2 つの翻訳方向でテストした。そして、それぞれの翻訳方向において、それぞれの単語を単独、そして簡単なセンテンスの中でテストした。英日のテストは 2 名の日本人、日英の翻訳テストは 2 名の英語母語話者によって採点された。これらの形式のテストを結びつけ、受験者、設問、採点者比較する為にラッシュ分析を行った。

Results. 日本語から英語への発表的翻訳の方が受容的翻訳よりも難しいという結果が得られたが、Webb(2008)の結果と異なり、この難易度の差に習熟度による変化はみられなかった。そして２つの翻訳方向は２つの異なる次元の知識ではなく、１次元における 2 つの因子であるという結果になった。さらに語彙を文脈の中で与えることは単独で与よりも難易度への影響が高かった。

Conclusions. L1 や L2 で提示された語彙に関する知識は 1次元的な語彙知識上にあるといえるが、その語彙がどのように、どんな文脈の中で提示されるかも重要な因子である。これにより、語彙知識が文脈から独立したものであるとする「特性」論への疑問が生じる。この研究の結果は語彙知識の「相互作用」論、すなわち文脈の中でどのように語彙が使われているかが重要とされることを支持する結果となった。

Background

Competing models of language proficiency (e.g. Bachman, 1995; Breiner-Sanders, Swender, & Terry, 2001, p. 8; Council of Europe, 2008; Hale, et al., 1988; Halliday & Matthiessen, 2004; Henning, 1991; Pike, 1979) incorporate vocabulary, but disagreements arise over issues such as the definition of "word" (Gardner, 2007), the contribution of different features to meaning potential (Halliday & Matthiessen, 2004; Nattinger, 1980), and how vocabulary knowledge can be measured (Chapelle, 1998), with Mobarg (1997, p. 223) arguing that "…people's views of the role of vocabulary in language learning…, teaching, and testing differ systematically according to how they view language theoretically.", with structuralist orientations favoring learning of words based on frequency counts, and lexical approaches favoring teaching of associated chains of words based on collocational analysis.

Cowie (1981, p.203), in an issue of *Applied Linguistics* devoted to lexicography, describes a "…marked and altogether welcome revival of interest in vocabulary teaching, which has gathered momentum throughout the 1970s.", and Mobarg (1997, p. 203) reports on a resurgent "lexis orientation" supporting the consensus that vocabulary is crucial to language learning. Meara (2002) cites the work of Palmer presented in 1936 as the origin of an ongoing "vocabulary manifesto", noting "The agenda they are following is one that was largely laid down by the pioneers of the

1920's….it is not obvious that the study of vocabulary acquisition has made huge strides since those days." (2002, p. 406).

Read (1997) identifies three issues in vocabulary testing; construct definition, whether vocabulary should be assessed in whole texts and discourse, and whether discrete tests of vocabulary are needed, with a fundamental construct related problem raised by the emphasis in vocabulary research on knowledge, not ability. Language testers have increasingly favored performance assessment (Chapelle & Angelis, 2008; Hughes, 2003; McNamara, 1996), but vocabulary research addresses competence, following Hymes' (1972) model. Read questions the testing of isolated words and competence rather than performance ability (Read, 1997, p. 303), echoing Oller's earlier view (1979, p. 189) that "…knowing a word is knowing how to use it in a meaningful context…. Does it make sense then to insist on testing word-knowledge independent of the constraints that govern the relationships between words in discourse?"

The importance of context was well recognized by the middle of the twentieth century, from the ideas of the anthropologist Malinowski, who developed the term "context of situation" in *The Problem of Meaning in Primitive Language* (1923, p. 6) to describe the immediate environment of communication, writing "…the meaning of a word must always be gathered, not from a passive contemplation of this word, but from an analysis of its functions, with reference to a given culture." (Malinowski, 1923, p.6) and that "neither a Word [*sic*] nor its Meaning [*sic*] has an independent and self-sufficient existence." (Malinowski, 1923, p.6) . The importance of context is reinforced by considering the enormous number of meanings frequent words have. Zipf (1935, 1949) long ago pointed out that the highest frequency words have the most meanings. Fries (1945) documented this for some of the most frequent words in English. In his investigation he found that the "total number of different meanings recorded and illustrated in the Oxford English Dictionary

for the first five hundred words of the Thorndike Word List is 14,070; and for the first thousand it is nearly 25,000" (p. 40). Nation (1983, p.11, cited in Carter, 1987) counted the meanings of the 850 words of Basic English for a total of 12,425 meanings, thus, comprehension of language must rely heavily on contextual clues to allow the correct meaning to be assigned to words.

Malinowski (1935, vol.2, cited by Halliday & Hasan, 1985) insisted that the whole idea of 'semiotics', the Greek word for 'sign' as a basis for linguistics was a mistake and a mistake that he had made himself: "I opposed civilized and scientific to primitive speech, and argued as if the theoretical uses of words in modern philosophic and scientific writing were completely detached from their pragmatic sources. This was an error, and a serious error at that." (1935, vol.2, p.58, cited by Halliday & Hasan, 1985). Halliday and Hasan, in using this quotation, show how these ideas have been further adopted and adapted in the field of linguistics, later writing "The sign has tended to be seen as an isolate, as a thing in itself, which exists first of all in and of itself before it comes to be related to other signs." (Malinowski, 1935, vol.2, p.3, cited by Halliday & Hasan, 1985), and suggests a change to this notion of how we might learn words, "I would wish to modify this definition of semiotics and say that, rather than considering it as the study of signs, I would like to consider it as the study of sign systems . . . that are in themselves not sets of individual things, but rather networks of relationships." (Malinowski, 1935, vol.2, p.4, cited by Halliday & Hasan, 1985), a central premise of systemic linguistics. Halliday and Hasan discuss how Firth (1935, in Halliday & Hasan, 1985, p. 8) and Hymes (1967) wanted to take Malinowski's ideas further and develop a new form of linguistic theory from which we can begin to develop a set of teaching principles and schemata for a more robust form of vocabulary acquisition. Halliday and Hasan describe Firth's beliefs that "all linguistics was the study of meaning and all meaning was function

in a context" (op. cit., p. 8) with Hymes proposing "a set of concepts for describing the context of situation" (1935, in Halliday & Hasan, 1985, p. 9) that Malinowski had introduced.

Thus, in contrast to the claims of researchers working within the "vocabulary manifesto" that vocabulary is under-emphasized (e.g. Folse, 2004; Judd, 1978; Nation, 1990), vocabulary researchers have ignored broader developments in applied linguistics and language testing, leading to validity concerns. Validation problems are apparent in work on "receptive" and "productive" vocabulary, with Melka citing Morgan and Oberdeck (1930) as early proponents of the hypothesis that language users have distinctive "receptive" and "productive" vocabularies, but this is often assumed rather than demonstrated by empirical evidence (Melka, 1997). Gass (1988, p. 94), discussing Teichroew (1982), reports that "It is clear that learners do not have 'two' distinct sets of vocabulary…", yet this assumption still seems implicit in a number of studies, including Crow (1986), Laufer (1998), Laufer and Paribakht (1998), Fan (2000), Mochida and Harrington (2006), Webb (2008), and Yan and Nicoladis (2009).

Crow (1986) views vocabulary acquisition as primarily learning new labels for pre-existing concepts, but, as no two words are exact synonyms, productive usage requires learning connotations. "Productive" knowledge is defined as needed "to use it while speaking or writing (productive channels); receptive knowledge is what one needs to know in order to understand a word while reading or listening (receptive channels)." (Crow, 1986, p. 242), and claims that "Obviously a much larger body of knowledge is required for the former than the latter." (Crow, 1986, p. 242), assuming that production requires precise control of the connotation of words, but that reception typically only requires "extracting the general meaning of a written passage, relatively vague denotative knowledge is usually sufficient." (Crow, 1986, p. 243). A "semantic field approach" is thus advocated (Crow & Quigley, 1985), with understanding of connotations and

learning of discrete unrelated items discouraged and words acquired in inter-related semantic networks.

Laufer (1998) investigated the relationship between different aspects of vocabulary knowledge by Israeli students with "passive" tasks requiring matching of words with definitions, "controlled active" tasks where target words are prompted, and "free active" tasks where subjects produce words of their own choosing, respectively using the Vocabulary Levels Test (VLT) (Nation, 1990), the Productive Vocabulary Levels Test (Laufer & Nation, 1999), and the Lexical Frequency Profile (Laufer & Nation, 1995). Statistically significant differences were found in growth of different aspects of vocabulary knowledge, "passive" knowledge was greater than "active" and the gap increased with proficiency, but correlations between the scores were unremarkable and the classical test theory (CTT) (Bachman, 1990; Brown, 2005; Henning, 1987) use of raw scores does not justify the claims made, as the comparisons required interval level data and test equating was not conducted. Additionally, "dimensions of vocabulary knowledge", "type of vocabulary knowledge", and "stage of learning" were not defined, and issues such as dimensionality and developmental stages are not addressed. The same battery of tests was used by Laufer and Paribakht (1998), but with Canadian and Israeli students, finding that "the 3 dimensions of vocabulary knowledge developed at different rates." (Laufer & Paribakht, 1998, p. 366).

Fan (2000) used similar "passive" and "controlled active" task types as Laufer (1998), to test "production" of words known "passively" by students within small groups of Hong Kong Polytechnic University students, finding that "reception" exceeding "production", but comparing raw scores between small groups taking different tests led to severe methodological flaws.

Webb (2008) noted shortcomings in Laufer (1998) and Laufer and Paribahkt (1998) regarding the comparison of different tests, and so tested written word translation from Japanese to

English and English to Japanese. However, each subject translated different words in each direction and the tests were not equated. Additionally the inclusion of large numbers of loan words written using the "katakana" syllabary such as "ガソリン", borrowed from the English "gasoline", and which can be transcribed mechanically into the English alphabet as *gasorin*, raises construct validity concerns given the relative ease of writing these words compared with words using "kanji" pictographs of Chinese origin.

Although, as Brown (2005) points out, in sample sizes of greater than 30 persons and items, a normal distribution of scores is expected, usually allowing interval level comparisons, in order to claim generalizable findings in the way that Webb (Webb, 2008) does, tests of normality must be conducted. In the case of Webb's data, mean person scores far exceeded the ideal 50% level (Brown, 2005; Henning, 1987; Hughes, 2003), meaning that the CTT assumption of normally distributed scores was almost certainly violated. Although differing ratios of "productive" and "receptive" knowledge at different levels were found, this violation of psychometric assumptions invalidates the conclusions, leaving the issue of "productive" versus "receptive" vocabulary development unanswered.

Such flouting of psychometric considerations raises serious concerns, as Douglas (2001, p. 442) notes, "whereas the concept of demonstrating validity and reliability has been integrated into how language testing research is conducted, SLA researchers have generally failed to recognize the need to demonstrate these qualities." Raw CTT scores, as used by Laufer (1998), Laufer and Paribakht (1998), and Webb (2008) do not provide the invariant measurement needed for additive comparisons, nor the interval level data needed for the *t*-tests and ANOVA used (Bond & Fox, 2007; Field, 2009).

Borsboom gives a scathing critique of use of raw scores without a psychometric model to guide interpretation, "in the ideal psychometric world, nobody could publish a test without at least… the outline of a psychometric model, and an attempt to substantiate that idea empirically." (2006, p. 430), and points out that "The interpretation of group differences on observed scores…depends on the invariance of measurement models…in practice, however, group differences are often simply evaluated through the examination of observed scores" (Borsboom, 2006, p. 427), so equal intervals in the measurement scale must show equal intervals in the construct being measured, and these intervals must be invariant across easy and difficult items and more and less proficient persons. If we wish to compare test scores in the way that Laufer (1998), Laufer and Paribakht (1998), and Webb (2008) do, then additivity is required, and the invariant interval level measures provided by models such as the Rasch model (Bond & Fox, 2007) are necessary.

Chapelle (1998) notes that "…when researchers use…summaries to make inferences beyond the actual performance, they are working within the domain of psychological measurement…" (Chapelle, 1998, p. 33), and describes three models of SLA; trait models, behavioral models, and interactionalist models. Trait theorists attribute performance to characteristics of persons, and minimize context so as to test lexical items discretely, arguing that random sampling allows generalization, while behaviorists hold that the context in which performance occurs is of interest, so aim to replicate the settings in which performance is to be predicted. Generalization to other contexts is not claimed, so vocabulary is presented and elicited in discourse settings, not as isolated discrete lexical items.

Reductionist trait views of language and acquisition, assuming modular sub-units that additively combine, has proved inadequate (Beckner, et al., 2009; Larsen-Freeman, 2006; Larsen-

Freeman & Cameron, 2008), illustrated by the debate over vocabulary items in the TOEFL test (Bachman, 1986; Chapelle & Angelis, 2008; Henning, 1991; Pike, 1979; Stansfield, 1986). Language testers have adopted interactionalist models, attributing performance to the interaction between traits and context (Chapelle & Angelis, 2008; McNamara, 1996, 2001, 2006; McNamara & Roever, 2006; Mislevy & Yin, 2009), and Schoenemann (2009) describes a complex adaptive interactionalist model of the co-evolution of biological and cultural aspects of language, but the "vocabulary manifesto" persists with trait models despite their wider rejection (Chapelle, 1998). Interactionalist accounts allow that the relationships between traits are not linear and deterministic, but complex and emergent (Beckner, et al., 2009; Larsen-Freeman & Cameron, 2008), so traits can only be defined in context, for example, vocabulary use in discourse. This view of context constraining linguistic choices is central to the systemic model of linguistics (Matthiessen, 2009), so vocabulary size must be described within field, tenor, and mode, not in isolation, as Chapelle explains (1998). Similarly, Charles (2000, p. 519) emphasizes the central place of context to meaning, "…the more similar two words are, the more similar their linguistic contexts are…as speakers learn the meaning of a word from encountering its natural linguistic contests, they acquire a schematic contextual representation of the linguistic contexts of a word."

However, as Chapelle (1998) points out, socioacademic communities implicitly support particular perspectives on construct validity through the use of particular types of tests, and tests developed and validated as research instruments can end up being used for other purposes, with construct validity ignored. Construct validity is raised by Melka, "though authors generally insist on a dichotomy between reception and production…it is quite impossible to find a clear and adequate definition of what is meant by reception and production." (1997, p. 84). This is discussed in terms of "degrees of knowledge", often broken into stages, giving sequences of acquisition, for

example, imitation, reproduction, comprehension, and production. Melka (1997, p. 98) argues that "None of these techniques refer to specifically receptive or productive knowledge, but rather to a single aspect of vocabulary knowledge (e.g. meaning) which in itself has little to do with reception or production." Acquisition of "productive" knowledge begins before "receptive" knowledge is complete, so the focus should be on measuring knowledge, with different tests targeting different facets of this, with "reception" and "production" merely performance manifestations of competence.

Aims

This research aims to replicate Webb's (2008) investigation of the effect of translation direction on difficulty, using Rasch measurement to provide interval level data and invariant measurement (Bond & Fox, 2007), and to compare difficulty of discrete word translation with the same items provided in sentential context. Four hypotheses were tested:

H1: Translation difficulty for the L1 to L2 direction is greater than that of the L2 to L1 direction.

H2: The gap between translation direction decreases with increasing person ability.

H3: Sentence context reduces translation difficulty compared with isolated words.

H4: L1 to L2 and L2 to L1 translation will follow a single unidimensional trait.

Sampling

Tests were administered to a convenience sampling of 260 learners in 9 classes with a mean TOEIC Bridge score of 111, a standard deviation of 16, and a range from 84 to 138, in the first semester of 2009.

Method

The 180 words tested by Webb (2008), comprising 60 each from the 701 to 1900, 1901 to 3400, and 3401 to 6600 frequency bands of the COBUILD dictionary were tested in four different ways: L2 to L1 and L1 to L2 translation of a target word presented in isolation, and L2 to L1 and L1 to L2 translation of a target word presented in a simple sentence written with the intention of providing meaningful, but non-defining, context, with non-target vocabulary restricted to words from the first 1000 band as reported by the *Vocabulary Profile* (Cobb, 2007) website. Subjects were tested on 15 words from each frequency band in each of the four tests, encountering each of the 180 target words in only one context and direction of translation.

As the target word list included words normally written using all three Japanese character sets, syllabetic "katakana" for non-Japanese loanwords, and pictographic "kanji" and syllabetic "hiragana" for Japanese words, students were instructed to write the Japanese words in the simpler katakana or hiragana, rather than the more difficult kanji. The extremely regular nature of these syllabaries and the Japanese phonotactic system means that being able to pronounce a Japanese word implies being able to write it syllabetically. With kanji pictographs, however, knowledge of the meaning and phonological form does not imply knowledge of the graphological form, making kanji writing ability of questionable relevance to knowledge of English vocabulary. Thus it was decided that hiragana writing of the Japanese words would be accepted. No practical alternative to normal spelling of English words could be devised. Although some students are probably familiar with phonetic spelling, this was deemed unreasonable, and recording and grading oral translation was hopelessly impractical.

Written responses were manually rated by four raters, two Japanese female and two native-speaker male teachers of English. The Japanese raters rated the L2 to L1 responses, and the native

speakers rated the L1 to L2 responses on a scale of 0 to 2, with native speakers rating "0" as unambiguously correct, "1" as minor errors judged unlikely to affect understanding, and "2" as unambiguously correct. Japanese raters used "2" to indicate the correct target word, "1" to indicate correct translation of a different sense of the English word, and "0" for an unambiguously incorrect response.

Two-faceted Rasch measurement using the Andrich rating scale model in the *Winsteps* software package (Linacre, 2007b) allowed concurrent equating and the additive measurement needed to compare the facets of persons and items on a common, invariant scale. Rasch measurement allows estimation of person and item measures despite missing responses, so responses coded as "1" can be recoded as missing to allow investigation of the effects of different rating systems, but this is beyond the scope of this preliminary report.

Results

The initial administration resulted in insufficient data for analysis of 90 items, so these items were eliminated from this working draft. For the remaining items, only responses where the two raters agreed were retained, with rater agreement of approximately 94.5%. Of the remaining 630 items, 146 were excessively easy or difficult for the sample of persons, resulting in large standard errors, and a further 16 of the remaining 484 items showed negative point-measure correlations, indicating that they were not measuring consistently with the other items in the test, and were thus removed pending full item analysis once the complete dataset is available. 27 more items were removed due to statistically significant misfit, leaving 236 persons and 441 items, with mean item measure specified as 100 and 1 logit scaled to 20. Finally items were identified where item difficulty for both Japanese-English and English-Japanese translation were measured, leaving

286 items of the original 720. Eight non-Japanese students were removed prior to analysis and 19 of the 252 remaining persons showed statistically significant fit problems, meaning that considerable "noise" was affecting the measurement of these persons. As the major facet of interest in this report is item difficulty, these misfitting persons were removed, leaving 232 of the original 260 persons.

Tables 1 and 2 show Rasch person reliability of .92 and item reliability of .95. Mean person and item measures are well matched, but there are few persons matched to the most difficult items, reflected in the standard deviation of item difficulty exceeding that of person ability.

Table 1

Summary of 232 measured persons

	Raw Score	Count	Model Measure	Model Error	Infit *MS*	Infit Zstd	Outfit *MS*	Outfit Zstd
M	66.7	62.5	101.79	3.88	1.01	.0	.93	-.1
SD	30.9	17.3	15.04	1.02	.22	1.1	.37	.8
Max.	162.0	93.0	141.44	13.13	1.83	2.8	2.56	2.7
Min.	2.0	7.0	47.63	2.87	.48	-3.3	.28	-1.9
Real Rmse 4.22 Adj.Sd 14.43 Separation 3.42 Person Reliability .92								
Standard Error Of Person Mean = .99								

Table 2

Summary of 286 measured items

	Raw Score	Count	Model Measure	Model Error	Infit *MS*	Infit Zstd	Outfit *MS*	Outfit Zstd
M	54.1	50.7	100.13	4.54	1.01	.0	.92	.0
SD	35.9	12.9	21.88	1.33	.22	1.3	.35	.9
Max.	134.0	71.0	145.68	9.26	1.68	2.8	2.49	2.8
Min.	4.0	16.0	53.10	2.75	.28	-9.9	.19	-5.4
Real Rmse 4.96 Adj.Sd 21.31 Separation 4.30 Item Reliability .95								
Standard Error Of Item Mean = 1.30								

Table 3 shows the results from paired samples t-tests on the item difficulties of the 143 pairs of L1-L2 and L2-L1 items, L1-L2 translation being approximately 1 logit, or 20 scaled units

more difficult, statistically significant at the $p < .001$ level. Thus, a person with a 50% expectation of success on L2-L1 translation has only a 27% expectation of success on L1-L2, supporting H1.

Table 3

Paired samples t-test of difficulty of translation direction

		Paired Differences							
					95% CI				Sig. (2-tailed)
		M	*SD*	*SEM*	*LL*	*UL*	*t*	*df*	
Pair 1	L2-L1 L1-L2	-20.49	17.24	1.44	-23.34	-17.64	-14.21	142	.000

Note. CI = confidence interval; *LL* = lower limit, *UL* = upper limit.

Figure 1 compares item difficulty of matched L1-L2 and L2-L1 translation pairs, with the dashed identity line crossing the x-axis at approximately 20, reflecting the greater mean difficulty of the L1-L2 direction, but many items fall outside the solid upper and lower 95% confidence bands, so difficulty of a word in one translation direction does not usefully predict difficulty in the other.

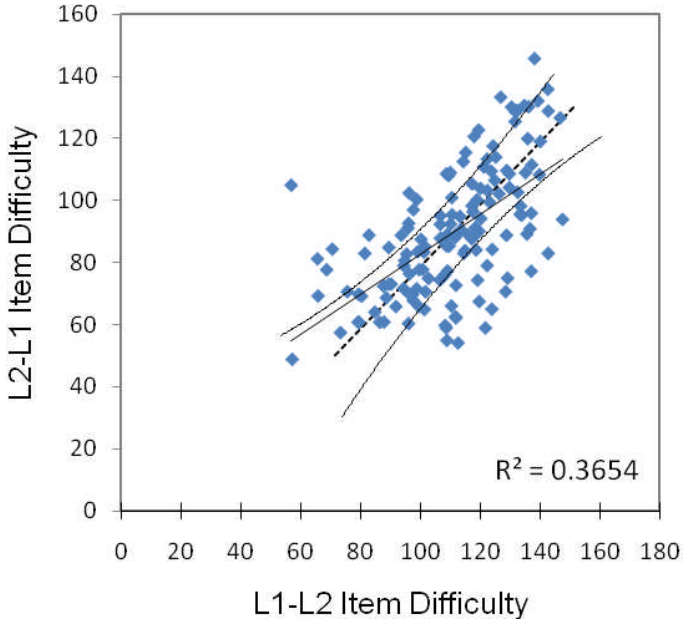

Figure 1. Item difficulty of L1-L2 versus L2-L1 translation direction. The difficulty of translation of words in the L1 to L2 direction is compared with the difficulty of translation in the L2 to L1 direction. The central dashed line shows the identity line. The central solid line shows the linear trendline. The upper and lower solid lines show the 95% confidence bands.

Comparisons of person ability confirm the average difference between translation direction. Person abilities were calculated independently for the two subsets of L1-L2 and L2-L1 items, with item difficulties anchored at values giving mean item difficulty of 100 for each sub-test. Table 4 shows the L1-L2 direction has resulted in person abilities approximately 20 points, or 1 logit, higher than the L2-L1 direction, statistically significant at the $p < .001$ level. What is of greater interest, however, is comparison of Figure 2, comparing person abilities for different translation directions, with Figure 1. Person ability is much more consistent than item difficulty, the magnitude of the differences is smaller, and the linear trendline more closely tracks the identity line. This greater consistency is reflected in the shared variance of 65% by the tests of persons, but only 37% by items. In short, the sample of persons is behaving relatively consistently between the two translation directions, but the sample of items is not.

Table 4
Paired samples t-test of person ability of translation direction

		Paired Differences							
					95% CI				Sig. (2-tailed)
		M	*SD*	*SEM*	*LL*	*UL*	*t*	*df*	
Pair 1	L1-L2 L2-L1	20.67	10.25	.69	19.32	22.02	30.18	223	.000

Note. CI = confidence interval; *LL* = lower limit, *UL* = upper limit.

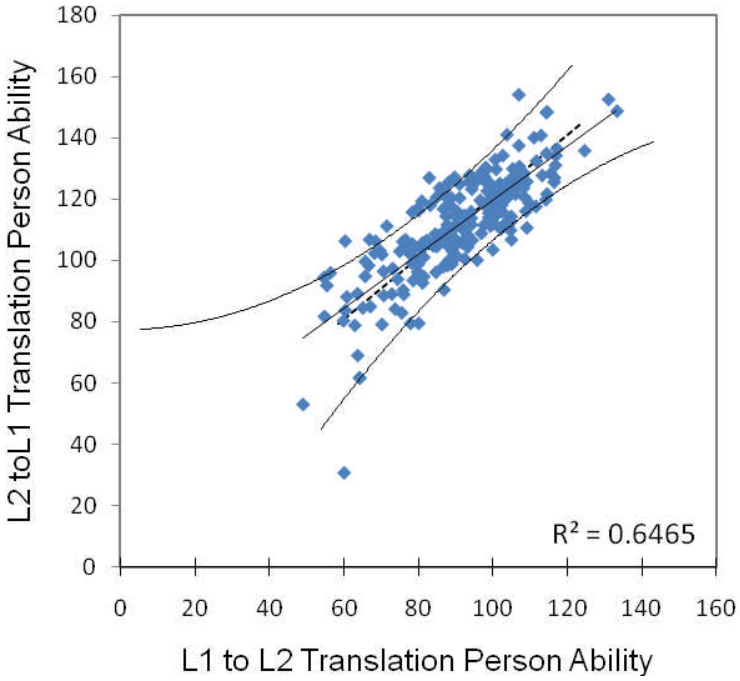

Figure 2. Person ability of L1-L2 versus L2-L1 translation direction. The ability of translation of words in the L1 to L2 direction is compared with the ability of translation in the L2 to L1 direction. The central dashed line shows the identity line. The central solid line shows the linear trendline. The upper and lower solid lines show the 95% confidence bands.

The two subtest scores were then added together and the top and bottom 5% of persons removed to reduce the influence of outlying results, leaving 201 persons who were divided into 3 groups, low (42 persons), medium (76 persons), and high (83 persons), each group having an approximately equal range of scores. 1-way ANOVA was used to analyze difference between group means, shown in Table 5. Although the low group shows a slightly higher difference than the high group, at less than .10 logits this is not substantively meaningful, and none of the differences

are statistically significant. Thus, L1-L2 translation is significantly and meaningfully more difficult, and this difference is consistent across ability levels. Thus H2 is rejected, and Webb's (Webb, 2008) finding that the gap decreased with increasing proficiency seems to be an artifact of methodological flaws arising from inappropriate use of raw scores.

Table 5

Multiple comparisons of differences in translation ability by level (Scheffe)

(I) Group	(J) Group	(I-J) Mean Difference	*SE*	Sig.	95% CI *LL*	95% CI *UL*
Low	Medium	.48	1.81	.966	-3.98	4.94
	High	1.84	1.78	.586	-2.55	6.24
Medium	Low	-.48	1.81	.966	-4.93	3.98
	High	1.38	1.49	.658	-2.32	5.05
High	Low	-1.84	1.78	.586	-6.24	2.55
	Medium	-1.38	1.49	.658	-5.05	2.32

Note. CI = confidence interval; *LL* = lower limit, *UL* = upper limit.

The item difficulty of the word and sentence cued items were compared, with person abilities anchored at values reported from the combined test but item difficulties allowed to float. The mean difficulties of the sentence cued and word cued items were 98.21 and 101.79 respectively, as shown in Table 6, the 3.58 unit difference being statistically significant to $p < .001$.

Table 6

Sentence versus word cued paired item difficulty

		Paired Differences							
					95% CI				Sig. (2-tailed)
		M	*SD*	*SEM*	*LL*	*UL*	*t*	*df*	
Pair 1	Sentence Word	-3.58	9.37	.78	-5.12	-2.05	-4.61	144	.000

Note. CI = confidence interval; *LL* = lower limit, *UL* = upper limit.

Figure 3 shows the scatterplot of word-cued and sentence-cued item difficulties, with the coefficient of determination of .79 reflecting a raw correlation of .89, increasing to .99 after

disattenuation. 18 items fall outside the 95% confidence band, somewhat more than expected from the sample of 145 items, suggesting the need for a detailed item analysis once the dataset is complete, but the linear trendline is close to the plot identity line. Although some items do show meaningful and significant differences in the word-cued and sentence-cued items, overall the two sub-tests perform very similarly overall and the mean difference of 3.58, or .18 logits means that a person with an expectation of success of 50% on an item of mean difficulty in the word cued test would have a 55% expectation of success on an item of mean difficulty in the sentence cued test. H3 is thus supported, sentential context results in reduced item difficulty.

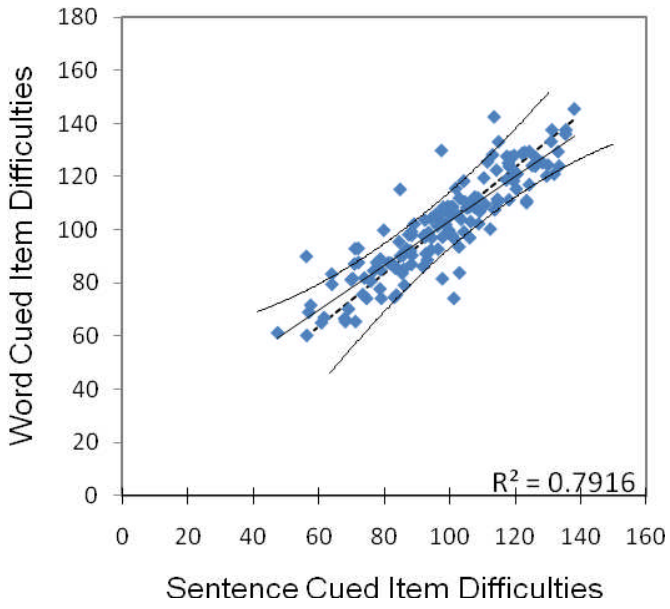

Figure 3. Item difficulty of word-cued versus sentence-cued translation. The difficulty of translation of words presented as isolated words and in a sentential context in both the L1 to L2 and L2 to L1 directions is compared. The central dashed line shows the identity line. The central solid line shows the linear trendline. The upper and lower solid lines show the 95% confidence bands.

Dimensionality of the combined test was investigated using *Winsteps* (Linacre, 2007b) principal components analysis of the Rasch residuals, allowing the size of the unidimensional

Rasch trait to be contrasted with the largest contrasting dimensions. Table 7 shows that the Rasch measures account for 48.2% of the variance in the data, with items accounting for 31.3% and the first contrast accounting for only 1.1% of the data. Thus the results are strongly unidimensional and H4 is supported; each translation direction measures a facet of unidimensional construct, with the differences in difficulty showing development steps along that dimension, contrary to Webb's (2008) unsupported assertion that they measure different dimensions.

Table 7

Dimensionality of the combined set of items

	Empirical		Modeled
	Eigenvalue	%	%
Total raw variance in observations	559.7	100.0%	100.0%
Raw variance explained by measures	269.7	48.2%	40.7%
Raw variance explained by persons	94.7	16.9%	14.3%
Raw Variance explained by items	175.0	31.3%	26.4%
Raw unexplained variance (total)	290.0	51.8%	59.3%
Unexplned variance in 1st contrast	5.9	1.1%	2.0%
Unexplned variance in 2nd contrast	5.2	.9%	1.8%
Unexplned variance in 3rd contrast	5.0	.9%	1.7%

The results show that, contrary to earlier claims, "productive" and "receptive" translation are not separate dimensions of knowledge. Although "productive" knowledge lags "receptive" knowledge, the 1 logit mean difference and variability of the gap means that "productive" knowledge begins to develop long before "receptive" knowledge is complete, supporting Melka's (1997) argument that these are not separate constructs, and casting doubt on Crow and Quigley's (1985) proscription against teaching productive use.

The "vocabulary manifesto" emphasis on testing discrete words isolated from context, ignoring Gass' (1988, pp. 95-96) concern that "…vocabulary instruction is often haphazard…in instances of specific lexical instruction, it is often the case that we deal with isolated words or, at best, words embedded within a single sentence.", conflicts with the finding that sentential context

significantly affects item difficulty, and also with the relative lack of stability in individual item measures between translation directions. A word does not have an inherent difficulty, but difficulty results from the interaction of multiple facets of items, persons, and contexts. An interactionalist view of vocabulary knowledge is thus necessary to address the likelihood of successful comprehension or production in different linguistic or discourse contexts, as is done in the broader field of language testing.

The validation of H3 which showed that a person with an expectation of success of 50% on an item of mean difficulty in the word cued test would have a 55% expectation of success on an item of mean difficulty in the sentence cued test has potentially important teaching implications, suggesting that providing context to help students better understand vocabulary and integrate it into their existing networks of knowledge is vital in planning, methodology, and curriculum. Halliday and Hasan exemplify the idea of context with "…we always have a good idea of what is coming next, so that we are seldom totally surprised. We may be partly surprised; but the surprise will always be within the framework of something that we knew was going to happen". (1985, p. 5), a phenomenon he views as the most important aspect of human communication.

It is this notion of a "framework" that perhaps best helps educators address the language learning strategies of our students. By introducing the key "frameworks" of communication in the classroom – how these frameworks are constructed and how the elements within them are related, when students experience these frameworks again, they will have some sense of what is going to happen and furthermore will gradually be able to experience or reconstruct these frameworks independently in a range of different contexts, be that through the framework of a sentence such as used in the tests constructed for this study, or the larger frameworks of a curriculum.

There is thus a body of theory and research upon which the teaching implications of H3 may be based, namely that of a "Text-based syllabus", following Halliday's definition of a text as combining "Field", what is happening, "Tenor", who is taking part, and "Mode", what part the language is playing, with "syllabus" defined by Yalden (1987 in Feez & Joyce, 1998) as a tool in which the teacher "can achieve a certain coincidence between the needs and aims of the learner, and the activities that will take place in the classroom.", echoing Lantolf's (2009) description of teachers intervening in an "ongoing flow", rather than simply presenting decontextualized facts unconnected to existing frameworks of knowledge.

Feez and Joyce (1998), further define a "Text-Based Syllabus" as having a view of language that "occurs as whole texts which are embedded in the social contexts in which they are used" and that people learn language through working with whole texts rather than through words that exist in and of themselves with reference to holistic models of content and methodology. The content of this syllabus must therefore be selected in relation to learner needs and the social contexts which learners wish to access.

Feez and Joyce explain that texts have lexical-grammatical features that allow native speakers to identify them, so highlighting these features rather than focusing exclusively on isolated words and phrases is necessary to help them develop the skills needed to first comprehend the texts and then construct their own texts through speaking and writing. This deconstruction of existing texts highlights to learners how individual words relate to a broader meaning. In this way students can use "context" to examine and "have a good idea of what is going to come next", thus formulating the meaning of new words and learning to utilize vocabulary to communicate more efficiently.

Butt *et al.* (2000) refer to the "text-based syllabus" as teaching "whole units of meaning" with five distinct elements, the most pertinent in this study being the "micro-elements of a text" or the "lexicogrammar" and how words are patterned within whole texts in context, and explain that "where students are not yet able to manage the whole text, the teacher fills in what the student is not yet able to do" (Butt et al., 2000, p. 10). In this way the students "experience what it is like to participate in the use of a whole unit of meaning" (Butt *et al.*, 2000, p.10), learn to identify the structures and lexico-grammars of a text, and finally create texts independently.

Thus the role of the facet of context in vocabulary knowledge will be a major avenue of future inquiry, so a reduced subset of the best functioning target words from this investigation will be used to develop tests presenting the words in the context of meaningful texts, the common items allowing the datasets to be linked and compared using MFRM, thus showing the contribution of different contexts to the difficulty of comprehending or producing L2 vocabulary, information which is essential for effective curriculum sequencing and task design.

Although the results presented here are statistically significant and the sizes of the differences are meaningful in everyday terms, many-faceted Rasch measurement (MFRM) (Linacre, 2007a; McNamara, 1996) of the complete dataset in 2010 will allow more detailed analysis of raters and other facets of test performance. MFRM allows the effect on performance of facets such as raters and item types to be measured on the same invariant scale as the facets of person ability and item difficulty. Another benefit of MFRM is that interactions between facets can be investigated, allowing the effect of "katakana" loanwords to be analyzed. Further tests were administered to four more classes in September 2009, and this complete dataset will be analyzed in February 2010, with definitive results expected in April 2010 for submission to *Studies in Second Language Acquisition* as a replication of Webb's (Webb, 2008) study.

References

Bachman, L. (1986). The test of English as a foreign language as a measure of communicative competence. In C. Stansfield (Ed.), *Toward communicative competence testing: Proceedings of the second TOEFL invitational conference*. Princeton: Educational Testing Service.

Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. (1995). What does language testing have to offer? In H. Brown & S. Gonzo (Eds.), *Readings on second language acquisition.* (pp. 415-447). Upper Saddle River: Prentice Hall Regents.

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., et al. (2009). Language is a complex adaptive system: Position paper. *Language Learning, 59,* 1-26.

Bond, T., & Fox, C. (2007). *Applying the Rasch model* (2 ed.). London: Lawrence Erlbaum Associates.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*(3), 425-440.

Breiner-Sanders, K., Swender, E., & Terry, R. (2001). ACTFL proficiency guidelines: Writing Retrieved 1 November, 2007, from http://www.actfl.org/files/public/writingguidelines.pdf

Brown, J. D. (2005). *Testing in language programs.* New York: McGraw-Hill.

Butt, D., Fahey, R., Feez, S., Spinks, S., & Yallop, C. (2000). *Using functional grammar: An explorer's guide* (2 ed.). Sydney: National Center for English Language Teaching and Research.

Carter, R. (1987). *Vocabulary: An applied linguistics perspective.* London: Allen & Unwin.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32-70). Cambridge: Cambridge University Press.

Chapelle, C. A., & Angelis, P. (2008). The evolution of the TOEFL. In C. A. Chapelle, M. K. Enright & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 27-54). New York: Routledge.

Charles, W. G. (2000). Contextual correlates of meaning. *Applied Psycholinguistics, 21*(04), 505-524.

Cobb, T. (2007). Web Vocabprofile Retrieved September 16, 2007, from http://www.lextutor.ca/vp/

Council of Europe (2008). *Common European framework of reference for languages: Learning, teaching, assessment.*

Cowie, A. P. (1981). Lexicography and its pedagogic applications: An introduction. *Applied Linguistics, II*(3), 203-206.

Crow, J. T. (1986). Receptive vocabulary acquisition for reading comprehension. *Modern Language Journal, 70*(3), 242.

Crow, J. T., & Quigley, J. R. (1985). A semantic field approach to passive vocabulary acquisition for reading comprehension. *TESOL Quarterly, 19*(3), 497-513.

Douglas, D. (2001). Performance consistency in second language acquisition and language testing research: a conceptual gap. *Second Language Research, 17*(4), 442-456.

Fan, M. (2000). How big Is the gap and how to narrow it? An investigation into the active and passive vocabulary knowledge of L2 learners. *RELC Journal, 31*(2), 105-119.

Feez, S., & Joyce, H. (1998). *Text-based syllabus design.* Sydney: National Center for English Language Teaching and Research.

Field, A. (2009). *Discovering statistics with SPSS*. London: Sage.

Folse, K. S. (2004). *Vocabulary myths.* Ann Arbor: The University of Michigan Press.

Fries, C. C. (1945). *Teaching and learning English as a foreign language*. Ann Arbor: University of Michigan Press.

Gardner, D. (2007). Validating the construct of *Word* in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics, 28*(2), 241-265.

Gass, S. M. (1988). Second language vocabulary acquisition. *Annual Review of Applied Linguistics, 9*(1), 92-106.

Hale, G. A., Stansfield, C., Rock, D. A., Hicks, M. M., Oller, J. W., Jr, & Butler, F. A. (1988). *Multiple-choice cloze items and the TOEFL® test.* Princeton: Educational Testing Service.

Halliday, M., & Hasan, R. (1985). *Language, context and text: Aspects of language in a social semiotic perspective*. Victoria: Deacon University Press.

Halliday, M., & Matthiessen, C. (2004). *An introduction to functional grammar*. London: Arnold.

Henning, G. (1987). *A guide to language testing.* Boston: Heinle & Heinle.

Henning, G. (1991). *A study of the effects of contextualization and familiarization on responses to the TOEFL® vocabulary test items.* Princeton: Educational Testing Service.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

Hymes, D. (1967). Models of the interaction between language and social setting. *Journal of Social Issues, 23*, 8-28.

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp. 269-293). Harmondsworth: Penguin.

Judd, E. L. (1978). Vocabulary teaching and TESOL: A need for reevaluation of existing assumptions. *TESOL Quarterly, 12*(1), 71-76.

Lantolf, J. P. (2009). *Open discussion with James Lantolf*. Paper presented at the JALT 2009 International Conference.

Larsen-Freeman, D. (2006). On the need for a new understanding of language and its development. *Journal of Applied Linguistics, 3*(3), 281-304.

Larsen-Freeman, D., & Cameron, L. (2008). *Complex systems and applied linguistics*. Oxford: Oxford.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics, 19*(2), 255-271.

Laufer, B., & Nation, I. S. P. (1999). A vocabulary-size test of controlledproductive ability. *Language Testing, 16*(1), 33-51.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*(3), 307-322.

Laufer, B., & Paribakht, T. S. (1998). The relationship between passive and active vocabularies: The effects of language learning context. *Language Learning, 48*(3).

Linacre, J. M. (2007a). Facets Retrieved 1 November, 2007, from http://www.winsteps.com/facets.htm

Linacre, J. M. (2007b). Winsteps Retrieved 1 November, 2007, from http://www.winsteps.com/

Malinowski, B. (1923). The problem of meaning in primitive languages. In J. Maybin (Ed.), *Language and literacy in social practice: A reader* (pp. 1-11). Clevedon: Multilingual Matters.

Matthiessen, C. M. I. M. (2009). Meaning in the making: Meaning potential emerging from acts of meaning. *Language Learning, 59*, 206-229.

Maybin, J. (Ed.). (1993). *Language and literacy in social practice: A reader*. Clevedon: Multilingual Matters.

McNamara, T. F. (1996). *Measuring second language performance*. Harlow: Pearson Education.

McNamara, T. F. (2001). Language assessment as social practice: challenges for research. *Language Testing, 18*(4), 333-349.

McNamara, T. F. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly, 3*(1), 31 - 51.

McNamara, T. F., & Roever, C. (2006). *Language testing: The social dimension*. Oxford: Blackwell.

Meara, P. (2002). The rediscovery of vocabulary. *Second Language Research, 18*(4), 393-407.

Melka, F. (1997). Receptive versus productive aspects of vocabulary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition, and pedagogy* (pp. 84-102). Cambridge: Cambridge University Press.

Mislevy, R. J., & Yin, C. (2009). If language is a complex adaptive system, what is language assessment? *Language Learning, 59*, 249-267.

Mobärg, M. (1997). Acquiring, teaching and testing vocabulary. *International Journal of Applied Linguistics, 7*(2), 201-222.

Mochida, K., & Harrington, M. (2006). The Yes/No test as a measure of receptive vocabulary knowledge. *Language Testing, 23*(1), 73-98.

Nation, P. (1990). *Teaching and learning vocabulary*. New York: Newbury House.

Nattinger, J. R. (1980). A lexical phrase grammar for ESL. *TESL Canada Journal, 14*(3), 337-344.

Oller, J. W. (1979). *Language tests at school: A pragmatic approach*. London: Longman.

Pike, L. W. (1979). *An evaluation of alternative item formats for testing English as a foreign language*. Princeton: Educational Testing Service.

Read, J. (1997). Vocabulary and testing. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary* (pp. 303-320). Cambridge: Cambridge University Press.

Schoenemann, P. T. (2009). Evolution of brain and language. *Language Learning, 59*, 162-186.

Stansfield, C. (1986). *Toward communicative competence testing: Proceedings of the second TOEFL invitational conference*. Princeton: Educational Testing Service.

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition, 30*(1), 79-95.

Yan, S., & Nicoladis, E. (2009). Finding le mot juste: Differences between bilingual and monolingual children's lexical access in comprehension and production. *Bilingualism: language and cognition, 12*(03), 323-335.

Zipf, G. K. (1935). *The psycho-biology of language: An introduction to dynamic philology*. Boston: Houghton Mifflin.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge: Addison Wesley.

[i] Parts of this report draw on an unpublished paper by J. Lake of Fukuoka Jo Gakuin University, used with permission.