EXPLOITING PSYCHOLINGUISTIC PREDICTORS TO DEVELOP THE VOCABULARY OVERCLAIMING IN ENGLISH (VOCE) MEASURE

by

Patrick Dubois

Honours B.Sc., University of Western Ontario, 1985

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Psychology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2015

Abstract

The overclaiming technique (OCT) measures both knowledge and the tendency to exaggerate knowledge. While proven useful in several domains, deeper understanding of the nature of overclaiming, especially the role of foils, is hampered by lack of standardization and volatility of items. Here I present three studies using English vocabulary as a proxy for general knowledge. Studies 1 and 2 focused on item selection for a 50-item instrument dubbed Vocabulary Overclaiming in English (VOCE). Study 3 confirmed test-retest and concurrent validity. Finally, we examined the pooled item set to search for systematic patterns of linguistic predictors of item performance. In sum, the use of psycholinguistic predictors helped systematize item selection in developing the VOCE, a general measure of overclaiming with commendable psychometric properties.

Preface

Although the overclaiming technique (OCT) had already been well-established, little, if any, work had been done exploring item validation or Foil (fake item) generation. Neither had anyone applied OCT to English vocabulary. It was my idea to use vocabulary as an ideal proxy for general knowledge and to exploit psycholinguistic measures to guide item selection.

Under the supervision of Dr. Delroy Paulhus, I selected (or generated) vocabulary overclaiming items for each survey, assembled and administered each study, and analyzed all the data.

This research was conducted with the permission of the University of British Columbia Office of Research Studies Behavioural Research Ethics Board, certificate numbers H14-03113 and H14-02773.

Table of Contents

List of Tables

List of Figures

List of Symbols

*: $p < .05$

**: $p < .01$

***: $p < .001$

Glossary

Reals: items referring to genuine, existing entities (people, places, events, etc.).

Foils: items that do not exist: Hence no-one could be familiar with them.

List of Abbreviations

A: Accuracy (= H – F)

B: Bias (= (H + F)/2)

ELP: English Lexicon Project

F: False-alarm rate; proportion of Foils claimed as being familiar.

H: Hit rate; proportion of Reals claimed as being familiar.

HSP: Human subject pool

OCT: Overclaiming technique

VOCE: Vocabulary overclaiming in English

Acknowledgements

I would like to thank Dr. Delroy L. Paulhus for his support and guidance in this entire program of research. Thanks also to research assistants Robin Haverstock, Vivian Tong, Jennifer Trick and Maggie Wei, and the enthusiasm and encouragement of my large MA cohort.

This research was supported by a $17,500 Joseph Armand Bombardier Canada Graduate Scholarship-Master's award from the Social Sciences and Humanities Research Council of Canada.

Finally, I would like to thank Anita DeLongis and Jeremy Biesanz for agreeing to serve on my committee.

Chapter 1: Introduction

In Errol Morris's 2014 biographical documentary about Donald Rumsfeld, *The Unknown Known*, besides famously distinguishing between "known knowns" and "unknown unknowns", Morris has Rumsfeld describe "unknown knowns", that is, "things that you think you know, that it turns out you did not" (Morris, 2014). This latter notion may be the essence of the phenomenon of overclaiming.

1.1 The Overclaiming Technique

The fact that people differ in their willingness to claim familiarity with non-existent items has been developed into a useful psychological assessment tool. As operationalized by Paulhus and colleagues (2003), the overclaiming technique (OCT) involves asking participants to rate their familiarity with various items, some of which do not exist. Those researchers took a signal-detection approach (Swets, Tanner Jr., & Birdsall, 1961), treating a participant's claims for genuine items (Reals) as signal and claims for fake items (Foils) as noise. The tendency to claim each item type can be indexed *hit rate* (proportion of Reals claimed, abbreviated as H) and *false-alarm rate* (proportion of Foils claimed, F). These indices can then be combined in two composite measures, *Accuracy* (difference between Reals and Foils claiming, H - F = A) and *Bias* (overall tendency to claim familiarity, (H + F)/2 = B). These straightforward composites, labeled 'common-sense' measures by Paulhus and Petrusic (2010), have been shown to be effectively equivalent to a variety of complex measures (e.g., *d'* and *c*) traditionally used in signal detection (Macmillan & Creelman, 1991). Throughout this paper I will use the abbreviations H, F, A and B for brevity.

1.1.1 History

The notion of evaluating respondents using fake items has a long history (Raubenheimer, 1925; Anderson, Warner, & Spencer, 1984; Phillips & Clancy, 1972; Stanovich & West, 1989). It wasn't until recently, however, that the notion was elaborated into a systematic instrument to tap both knowledge accuracy and knowledge exaggeration. An up-to-date review is now available (Paulhus, 2011).

One motivation for developing the OCT was the need for a better measure of self-enhancement (Paulhus & Vazire, 2011). By asking people about their desirable qualities, self-report measures (e.g., social desirability scales) entail a problematic confound: Some people self-enhance but others actually do have the desirable qualities that they claim. Hence, genuine character cannot easily be separated from reported levels (e.g., Block, 1965; McCrae & Costa, 1983). As a behavioral measure, OCT overcomes that drawback: Respondents have to make specific memory claims about fake items.

Other defensible techniques require a comparison of self-reports with objective external criteria (e.g., John & Robins, 1994). However, the OCT is far more efficient by incorporating both self-report and criterion in a single questionnaire format. A number of critiques have been published over the last few years. One is that OCT Accuracy is not based on a solid objective criterion for knowledge (Ackerman & Ellingsen, 2014). Another is that OCT Bias taps openness to experience rather than self-enhancement (de Vries, in press). As a whole, however, the full body of research supports the utility of both OCT indices (Paulhus, 2011).

1.2 Item Content

The OCT was intended to serve as a methodological framework rather than a specific inventory with a fixed set of items. The original overclaiming questionnaire (OCQ) (Paulhus &

Bruce, 1990) included only academic content in 10 categories of 15 items each (e.g., science, law, philosophy, history, literature, language). The item set (primarily from Hirsch Jr., Kett, & Trefil, 1988) was selected to circumscribe the minimal cultural literacy of an educated American.

Subsequent studies with the academic OCQ found that Accuracy correlated with verbal IQ scores in the .40 - .60 range (Paulhus & Harms, 2004) and Bias correlated moderately (.25 - .38) with trait self-enhancement measures such as narcissism and self-deceptive enhancement (Paulhus et al., 2003).

A longer non-academic OCQ, with 10 items each in 25 domains relevant to less educated samples (fashion, sports, world leaders, etc.) revealed the importance of personal investment in a knowledge domain (Paulhus, Nathanson & Williams, 2005). Narcissism, for example, correlated with Bias only in domains that the respondent valued (consistent with Ackerman's (2000) notion that people are not invested in topics irrelevant to their identities). Nonetheless, higher IQ scores were associated with Accuracy for almost all lay topics, regardless of interest -- with the notable exceptions of professional wrestling and monster trucks.

These (and other) studies demonstrated the resilience of the technique, leading other researchers to implement, sometimes with little concern for item selection.

1.2.1 Hazards in Item Selection

Other researchers have developed more haphazard overclaiming measures. Randall and Fernandes (1991), for example, employed an overclaiming measure covering 10 popular culture domains of 5 items each, 2 of which were Foils. Their overall claim rate was only 9%, with 30% of respondents not endorsing any items. Although they found significant correlations with social desirability scales (.18*** with Marlowe-Crowne, .13* with BIDR (overall), .14** with SDE,

.11* with IM), null effects with other measures may be due to low claim rates and possible floor effects (no indication of variance was given).

In a recent followup, Joseph, Berry, & Deshpande, (2008) also used the Randall- Fernandes scale. Although they found a coherent pattern of associations, two of their foils (taken from the 1991 study) show up in a Google search as referring to genuine entities originating well before the 2009 study took place. In other words, Foils can become Reals, particularly when considering popular culture, a phenomenon I have noticed in examining other overclaiming scales. Might their results have been stronger with more careful item selection?

In examining the general public's mental health literacy, Swami, Papanicolaou, & Furnham, (2011) included an overclaiming measure with items related to mental health, with Reals such as dyslexia, schizophrenia and post-traumatic stress disorder. By far the highest claim rate (larger than those of all the other five Foils combined) was "Multiple Identity Syndrome". Because that term has been used legitimately in print, it hardly qualifies as non-existent. The average claim rate of the other Foils was a mere .09, each with standard deviations exceeding twice their means, suggesting a strong floor effect. (Later in this paper I will show how that can distort results). Remarkably, their Accuracy and Bias measures still had the strongest associations (compared to other big five and psychiatric skepticism measures) with measures of self-rated intelligence and knowledge of psychiatry. However, there was little distinction between Accuracy and Bias: They both predicted equally well with an inter-correlation of .60***. Apparently, weak performance of the Foils limited the distinction between Accuracy and Bias as independent predictors.

1.2.1.1 Item Issues

Applications of OCT such as those above point to a need for better understanding of item selection, particularly item generation for Foils. Other applications of OCT that yield null results may be misleading, in that they may be more a product of inefficient item sets than a property of overclaiming itself.

1.2.1.2 Item Volatility

I use the term 'volatile' to refer to items whose status as foils are subject to temporal change. For example, may change if the test were to become popular or is repeatedly used in the same context. Although it does not affect overclaiming accuracy (Paulhus et al., 2003), warning participants about the inclusion of foils has a number of side effects. It reduces foil claiming and undermines its utility as an individual difference measure. One insidious effect is the 'false fame' impact (Williams et al., 2002): Over time, Foils seem more real just because they have been observed in a previous study. Even more destructive, public exposure of fake items could lead to irrelevant individual differences in awareness of foils.

As shown above, the non-existence of a Foil can change at almost any time, particularly when dealing with popular culture. Although less dramatic, the relevance of a Real can change with time: Familiarity with "Watergate" may have been an indicator of political savvy in the 80s or 90s, but today may be more an indicator of age.

1.3 Recent Advances

Several recent advances highlight the importance of measuring Bias along with Accuracy in overclaiming research. One is the application of OCT to scholastic performance (Paulhus & Dubois, 2014). First, we showed that OCT Accuracy was comparable to multiple-choice tests and more efficient than short-answer questions in assessing undergraduate knowledge of

psychology. Moreover, they were comparable in reliability and predictive validity. Most important, we showed that OCT Bias (i.e., Foil claiming) provided an additional, incremental predictor of final grade, even controlling for mid-term mark or other knowledge measures (Paulhus & Dubois, 2014). The independent contribution of Bias highlights the need for strong, reliable Foil items in order to maximize potential predictive value.

Another advance emerged from the application of OCT to consumer research (Roeder & Paulhus, 2010). The need for Foils in consumer surveys is self-evident when asking respondents about their familiarity with new products. Yet many such survey include no Foils. When Foils were included in a series of large-sample survey of consumer products, a strong Bias factor emerged independently of Accuracy. The authors concluded that consumer surveys are pointless without taking into consideration individual differences in familiarity exaggeration (i.e., OCT Bias).

Finally, there is growing evidence that the act of overclaiming has a physiological substrate. One cognitive neuroscience study (using a version of the academic OCQ) found that transcranial magnetic stimulation (TMS) of the medial prefrontal cortex (MPFC) significantly reduced Foil claiming and improved Accuracy while also reducing reaction time. The implication is that social monitoring / reflection was mediating overclaiming behavior (Amati, Oh, Kwan, Jordan, & Keenan, 2010). Similar results emerged from a separate research laboratory (Beer & Hughes, 2012). These intriguing findings support the need for more investigation of overclaiming as a deeper neuroscientific phenomenon, not simply a covariate when examining other behaviors.

1.4 Research Goals

Earlier approaches to overclaiming relied on generalizing exaggeration across several domains of knowledge. That approach raises a multitude of individual difference confounds:

That is, respondents differ dramatically with respect to domains in which they are invested or knowledgeable. It also introduces noise for low-interest domains, and the challenge of selecting optimal Reals and Foils for each domain, since practicality permits only a few. For Foils, this raises the intractable question of what domain does a non-existent item belong to? For Reals, especially with popular culture domains, how does one choose appropriate level of difficulty? Indeed, the ad hoc nature of item selection (especially for Foils) has been termed by Paulhus as 'the shotgun approach" (personal communication).

For these reasons, I turned to the single content domain – English language vocabulary – a knowledge area with substantial documentation, expertise and with substantial supporting research. Despite being only a single knowledge domain, "vocabulary knowledge correlates highly with performance on more general measures of intelligence and is commonly viewed as a proxy for IQ" (Marchman & Fernald, 2008). Although all languages evolve, a substantial portion of vocabulary remains accessible across generations, providing more stability than most academic domains and virtually any popular domain. The choice of vocabulary also allows for tapping into the wealth of psycholinguistics research that may be helpful in optimizing item selection.

The main goals of this research program were (1) to develop an English vocabulary overclaiming measure with the acronym VOCE), (2) to explore what factors lead to better item performance, and (3) to use item performance predictors as a way to generate new items sets with predictable outcomes.

1.5 Psycholinguistic Analysis

Given that the knowledge domain for VOCE is English vocabulary, a wealth of information from existing psycholinguistic research is available to guide development of this instrument.

Beyond relying on *post hoc* analysis of item behavior from empirical survey data, the possibility of developing *a priori* techniques for item selection would make research more efficient and facilitate development of novel item sets.

1.5.1 Linguistic Resources

1.5.1.1 Vocabulary.com

Our source for vocabulary words to evaluate came from www.vocabulary.com, a site that advertises its computer-assisted language learning service that uses Item Response Theory to create "the fastest and most efficient way to master new words" (Zimmer, n.d.). They post their "Top 1000" vocabulary words "most likely to appear on the SAT, ACT, GRE, and ToEFL", in order of difficulty. Because the latter sources suited our student sample study population, I downloaded those words and used their rank (1-1000) as a Difficulty measure.

1.5.1.2 Wuggy

Wuggy is the name of a special software application designed to be a multi-lingual "pseudoword generator particularly geared towards making nonwords for psycholinguistic experiments" (Keuleers & Brysbaert, 2010) (http://crr.ugent.be/programs-data/wuggy). Wuggy generates several possible pseudowords from any given actual word, following user-selectable lexical constraints such as length (number of letters), subsyllabic element length (e.g. "tr-ee" has subsyllabic elements of length 2 and 2, "t-ee" has lengths 1 and 2), transition frequency (number of legitimate words in which two specific subsyllabic elements occur in the same sequence), and number of subsyllabic elements to match. Thus one uses genuine words as templates from which to generate lexically similar nonwords. With the multiple nonwords generated, Wuggy lists statistics such as OLD20 (see below) and deviation in transition frequency relative to the source genuine word.

1.5.1.3 WordGen

WordGen is another special software for generating nonwords satisfying different combinations of lexical constraints (Duyck, Desmet, Verbeke, & Brysbaert, 2004) (http://www.wouterduyck.be/?page_id=29). Unlike Wuggy, it generates words based on lexical parameters alone, not from a prototype genuine word.

1.5.1.4 English Lexicon Project

The main resource we used for psycholinguistic data was the English Lexicon Project (ELP; http://elexicon.wustl.edu/) which provides free online access to a large set of lexical characteristics and behavioral data on a corpus of 40,481 words and 40,481 nonwords (Balota et al., 2007). The behavioral data centers on the lexical decision task procedure, which involves timing subjects' classification of words and nonwords. The online database provides several lexical measures for each genuine word, and can provide some measures for an arbitrary list of nonwords.

While there are several psycholinguistic measures we explored, I will introduce here only those that we found had some relevance to our research.

1.5.1.5 Google

Regardless of how Foils are generated, no software or human can guarantee a "nonword" is actually not used as a word by somebody, somewhere. For this issue, there is no better resource than the omniscient Google search engine (www.google.ca). All Foils were tested by submitting them to Google to see if they had any use as slang, proper names or other non-standard usage. A surprising number of supposed "nonwords" had some usable meaning, according to Google, particularly short words. To pass the Google test, search results had to say "Did you mean …", or otherwise indicate that Google was stretching to find results. If results were something other

than the given word (Google hyphenated it, respelled or otherwise distorted it) or if results indicated something not viable as an English word (an unusual last name, an exotic foreign word, someone's bad spelling, etc.) it was considered viable as a nonword.

1.6 Reporting Conventions

Throughout this paper, the term Reals refers to genuine English word items and the term Foils refers to non-English word items. The letter H (for hit rate) will represent per-subject proportion of Reals claimed and F (for false-alarm rate) will represent per-subject proportion of Foils claimed.

As noted earlier, I will use the commonsense indicators of Accuracy (difference between Reals and Foils claiming, H - F = A) and Bias (overall tendency to claim familiarity, (H + F)/2 = B). All these statistics are scaled to be in the range 0 to 1, although A can dip below zero due to chance. Although negative Accuracy holds some theoretical interest (Roeder & Paulhus, 2010), those issues are not discussed here, and instances of negative A are treated as noise.

Since much of this work is exploratory, to remain conservative, all t-tests do not assume equal variance and the Welch (or Satterthwaite) approximation to the degrees of freedom is used.

All estimates are followed by 95% confidence intervals in square brackets, and asterisks are used to indicate p values: * = $p < .05$, ** = $p < .01$, *** = $p < .001$.

Chapter 2: Study 1

2.1 Item Selection

2.1.1 Reals

We began by selecting 25 real words from the site Vocabulary.com Top 1000 list described above: 15 were randomly chosen from the lower 100 (labelled "Easy" words) and 10 from the upper 100 ("Hard" words).

2.1.2 Foils

As a control, we attempted to maximize the lexical similarity of Foils and Reals, following the simple principle articulated Paulhus et al. (2003): "foils were created to appear to be plausible members of the same category". Given that we had only the one knowledge category, namely, English vocabulary, we were able adjust our similarity criteria.

For this purpose, we used the Wuggy software with maximal constraints to create Foils as lexically similar as possible to our selected Reals, in length, number of syllables, and unusualness of spelling. From the multiple nonwords Wuggy generated for each Real, the least unusual one (based on lexical measures) was chosen.

2.2 Method

2.2.1 Participants and Procedure

To get a large sample, we took advantage of an existing online survey used to prescreen participants for the human subjects pool (HSP) administered by the psychology department at UBC. Undergraduate students are incentivized by course credit to participate in the HSP and take the prescreen survey, which includes a variety of questions relevant to various researchers, and some basic demographic questions. These surveys are taken online at a time and place of the

student's choosing. The initial sample totaled 2922. Mean age was 20.2 with 71 percent female. More details about the sample are provided in Table 1.

2.2.2 Measures

The preliminary VOCE involved 50 items; 25 Reals and 25 Foils. These were presented, in a randomized order, after a stem of "Please rate your familiarity with the words below. This is just a general survey of UBC students, not an evaluation of your knowledge". The two response options were "Never heard of it" and "Heard of it somewhere". Responding with the latter option was taken as claiming some knowledge of the item.

The overall prescreen also included some demographic items: gender, ethnicity, age, and Native language. For analysis, we simplified these variables (other than Age) to binary: Female (1) vs. Male (0); Caucasian (1) vs. non-Caucasian (0); Asian (East Asian, Southeast Asian; 1) vs. non-Asian (0); native English spoken by age 4 (1), vs. non-native English (0). (See Appendix A for more detail.) To test for cultural differences we also isolated two subsets that should be polarized on the independent / interdependent (individualist / collectivist) cultural spectrum: Specifically, native English-speaking Caucasians, and native Japanese-speaking Japanese. Some authorities suggest that, compared to collectivistic cultures individualistic cultures promote self-enhancement tendencies (Heine & Lehman, 1997), which overclaiming is presumed to capture.

2.3 Results

Each item on the prescreen had a "Decline to answer" option, which would result in missing data. We selected only those respondents with no missing responses for the VOCE items. In general, if all VOCE items were answered, almost all of the other items were answered (and those with missing VOCE items often skipped several other items). Nonetheless, there remained

insignificant numbers of missing data on other items, which was not deemed problematic because overall N was so large. In total, 2922 usable surveys were collected.

Reporting of results are grouped into two sections: the behavior of VOCE items (Internal), and how total VOCE scores compared with other measures (External).

2.3.1 Internal Analyses

2.3.1.1 Distributions of H and F

Recall that H is the hit-rate of a respondent, the proportion of Reals they claim to know, and F is the false-alarm rate, the proportion of Foils they claim to know. The distributions of H and F indicate the appropriateness of the items for the sample. Ideally, distributions approximate centered normal curves with very few scores hitting the limits of the measure.

Unfortunately, the initial sample showed ceiling and floor effects for H (M = .88, SD = .11) and F (M = .10, SD = .14) respectively (see Figure 1). 35% of respondents claimed no Foils at all.

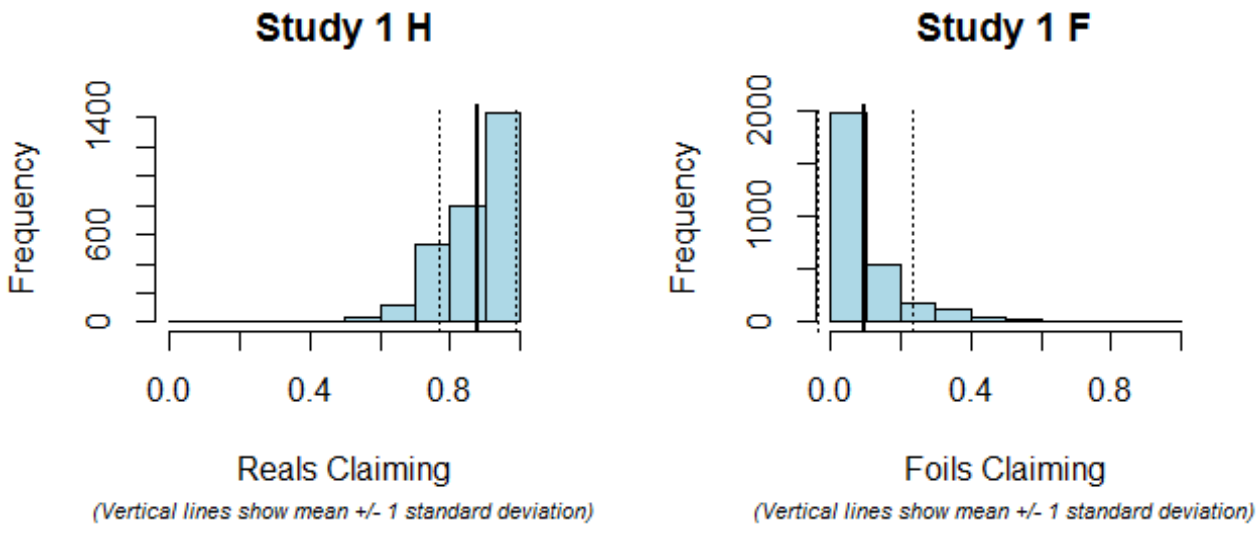


Figure 1. Study 1 distributions of per-subject claiming of Reals and Foils.

An examination of claim rates for individual items (Figure 2) confirms the pattern: Most Reals were claimed by nearly everybody, while Foils were typically claimed rarely.

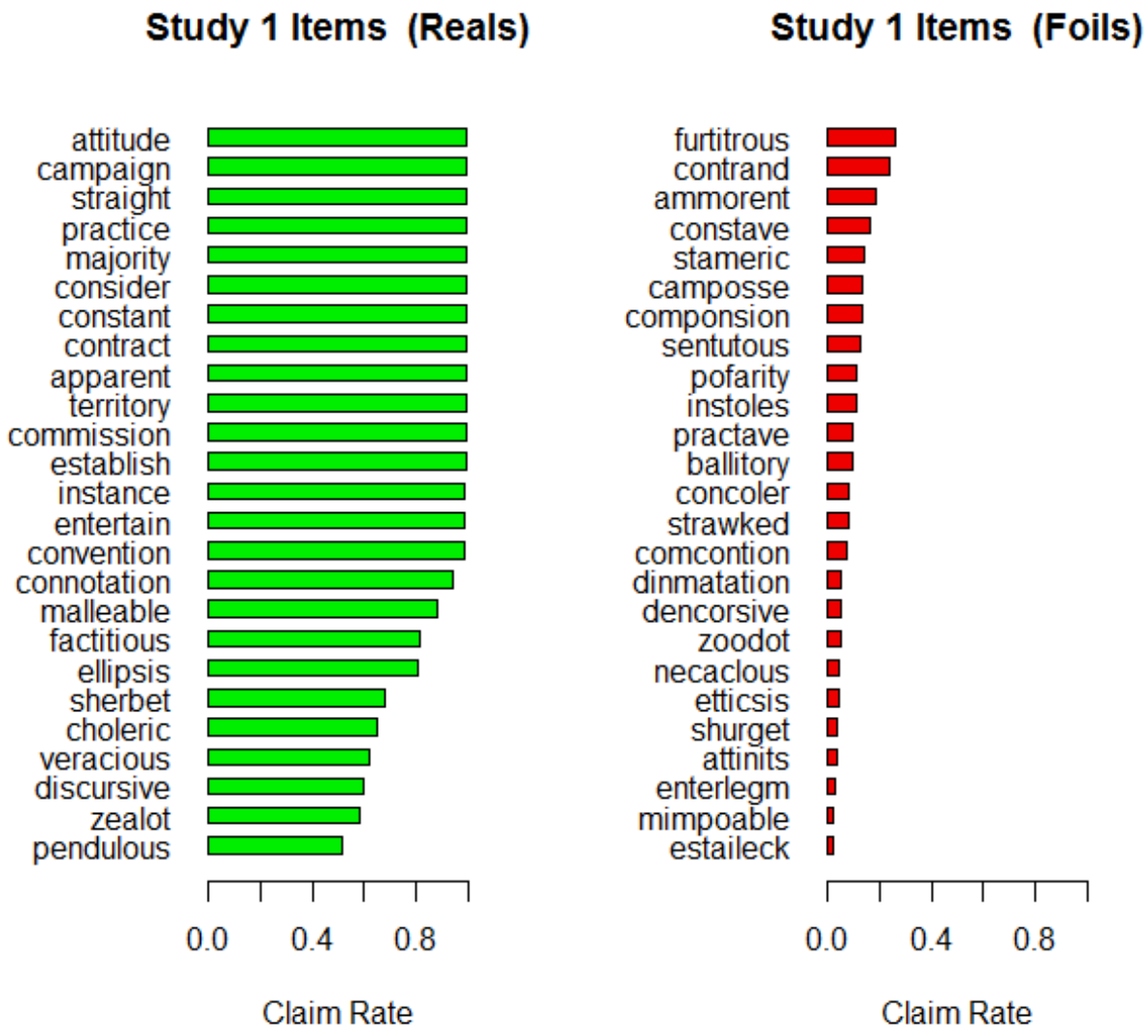


Figure 2. Study 1 Real and Foil items ranked by claim rates.

2.3.1.2 Item Subset Selection

To analyze how the survey might have performed without ceiling and floor effects, 20 non-extreme items were selected: 10 Reals with means below 95% (*choleric, connotation, discursive,*

ellipsis, factitious, malleable, pendulous, sherbet, veracious, zealot), and 10 Foils with means above 11% (*ammorent, camposse, componsion, constave, contrand, furtitrous, instoles, pofarity, sentutous, stameric*). I refer to these as the "Good" subset.

Using only the Good subset smooths the distribution for both Reals (M = .71, SD = .23) and Foils (M = .16, SD = .20), showing greater variance than the whole set, despite being less than half the number of items, with noticeably less ceiling effect for Reals (Figure 3).

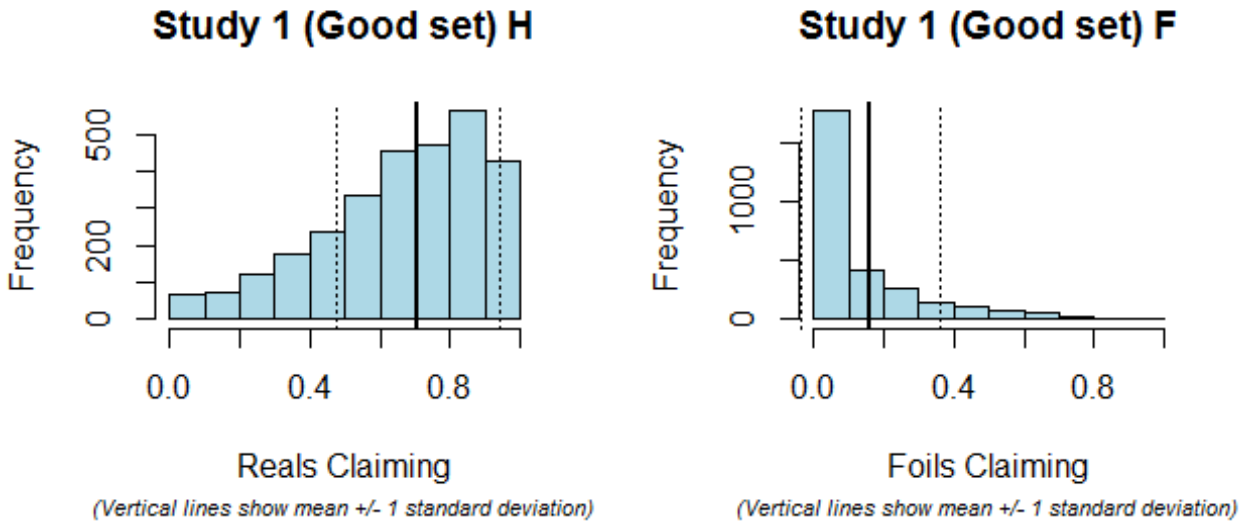


Figure 3. Study 1 Good subset distributions of per-subject claiming of Reals and Foils.

2.3.1.3 Distributions of A and B

Since participants on average claimed roughly half the items (Bias *M* = .49, *SD* = .09), a skeptical interpretation might be random answering, but that is belied by a negatively skewed Accuracy distribution, showing that respondents clearly distinguished Reals from Foils (Accuracy *M* = .78, *SD* = .16) (Figure 4).

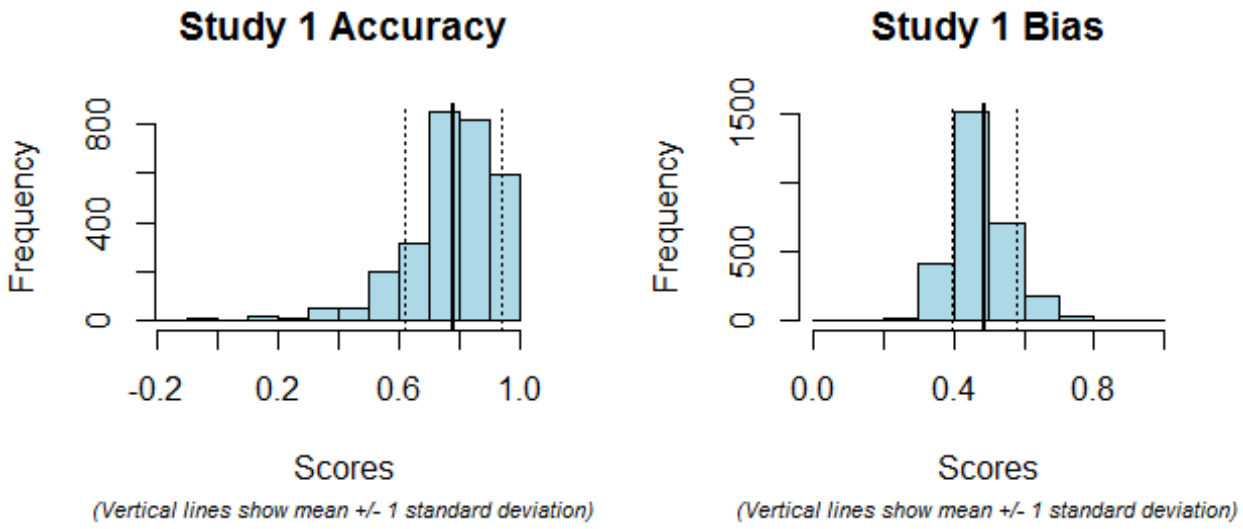


Figure 4. Study 1 distributions of per-subject Accuracy and Bias.

With the Good subset (Figure 5), variance is noticeably increased (Accuracy M = .55, SD = .27, Bias M = .44, *SD* = .17), again showing the importance of avoiding ceiling or floor effects.

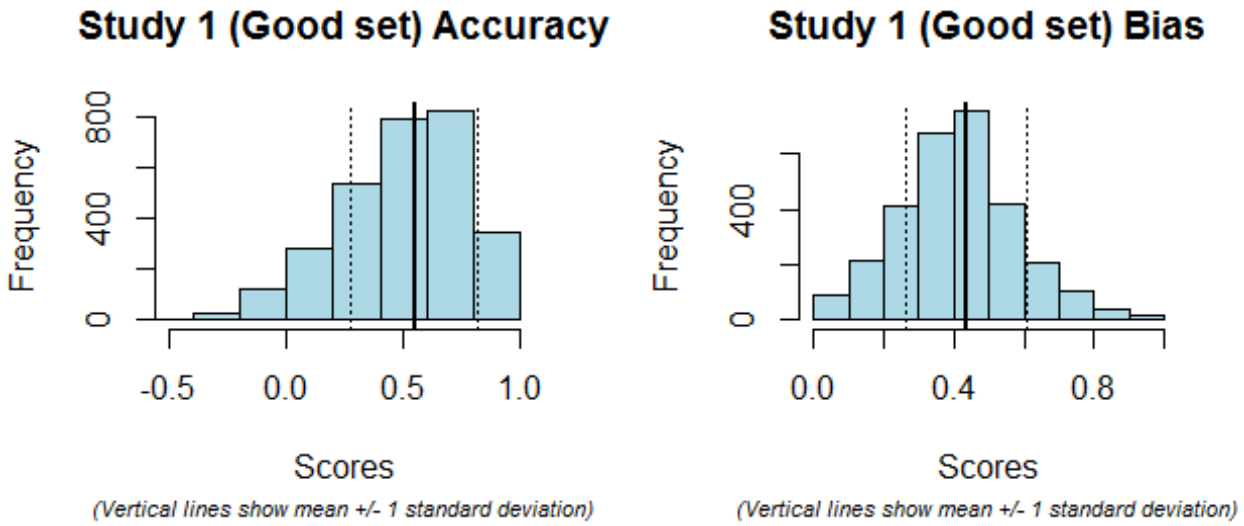


Figure 5. Study 1 Good subset distributions of A and B.

2.3.1.4 Correlations

H and F correlate positively (r(2920) = .14*** [.10, .17]), indicating that as one claims more Reals, one claims more Foils. On the other hand, A and B are negatively correlated (r(2920) = -.23*** [-.27, -.20]), suggesting that more knowledgeable individuals claim less.

With the smaller but more varied Good subset, the correlation between H and F increases to r(2920) = .23*** [.20, .27], while the relationship between A and B reverses to r(2920) = .17*** [.13, .20], suggesting that ceiling effects were indeed altering the behavior of the instrument.

2.3.1.5 Item Performance

For the purposes of this paper, the most straightforward measure of the value of an item is its item-total correlation with overall Accuracy (IT.A): Unless otherwise noted, that index will operationalize item performance here. Because Foils should, by definition, correlate negatively with Accuracy, their performance is measured as -IT.A. For easier comparison, the IT.A of Foils has been reversed, so that a negative performance for either Reals or Foils suggests an inappropriate item.

Neither claim rates nor variance predicted performance for either Reals or Foils. This finding is fortunate because it suggests that a) item variance is not artificially restricted by having binary variables (for which variance is a function of the mean), and b) a wide range of discriminability is available, i.e. frequently or rarely claimed items may both contribute to performance. However, the power to test this is limited by the small number of items.

However, it is noteworthy that performance for Foils (M = .37, SD = .06) was significantly higher, on average, than for Reals (M = .27, SD = .06), with a difference of -.10*** [-.13, -.07], t(47.90) = -6.10; d = 1.73. Figure 6 shows dispersion of claiming and performance for Reals and Foils.

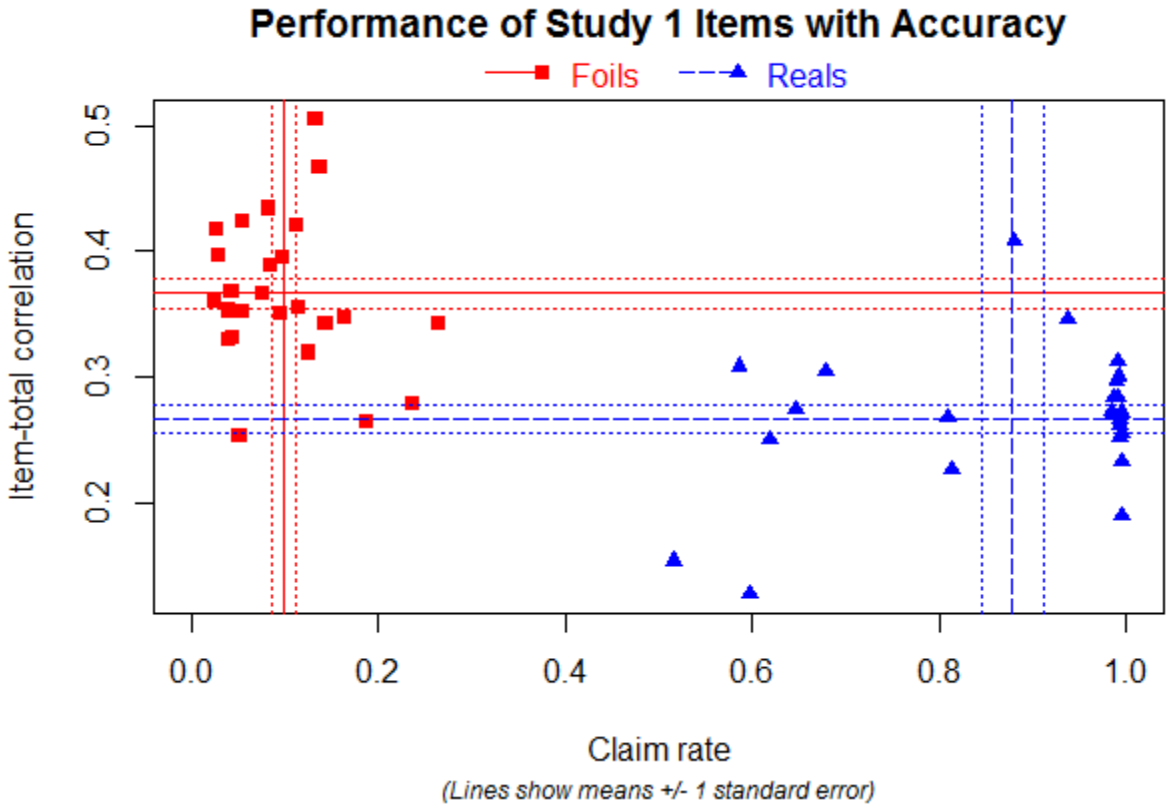


Figure 6. Study 1 claim rates vs performance.

Contrast that with the dispersion for the Good item subset in Figure 7, where Reals performance (M = .38, SD = .06) is now significantly higher than that of Foils (M = .30, SD = .09), by -.09* [-.16, -.02], t(16.14) = -2.68; d = 1.20.

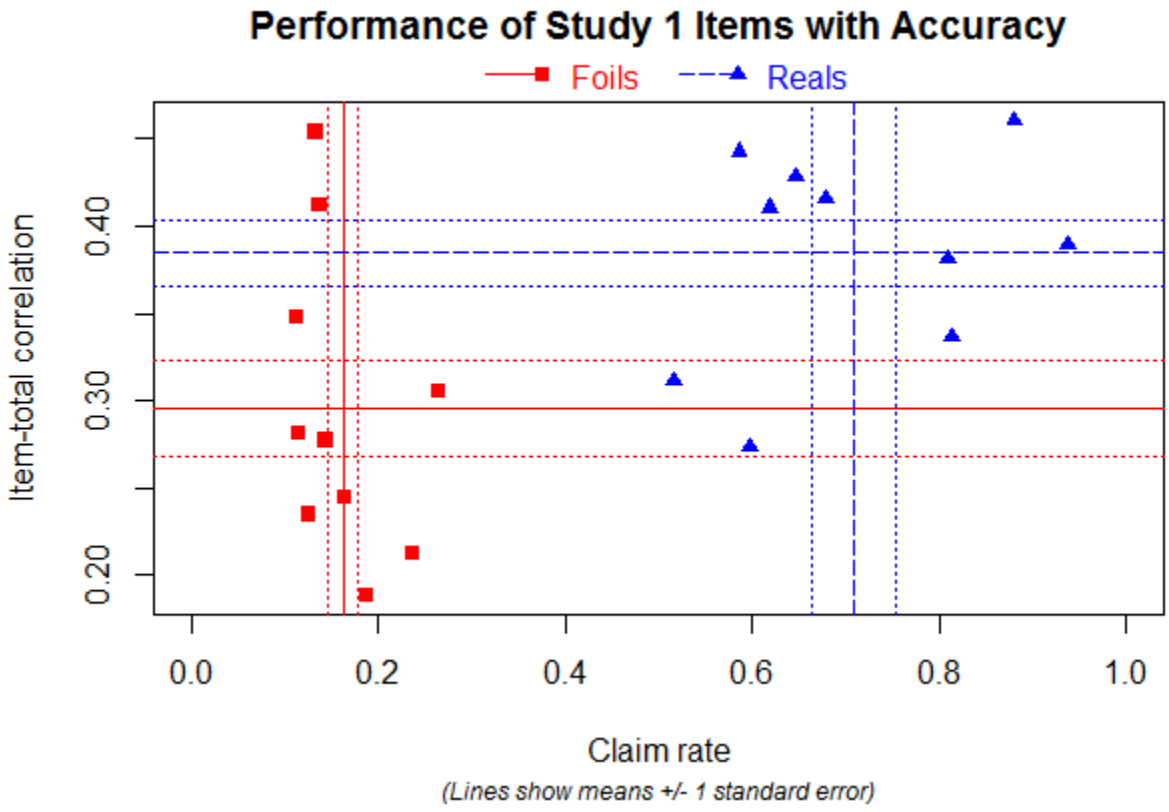


Figure 7 Study 1 Good subset claim rates vs performance.

2.3.1.6 Item pairing

Foils were generated to be as lexically similar to Reals as possible. To assess if this was a useful approach, we examined correlation of means and performance for the pairs.

Mean claim rates for Real-Foil pairs moved in opposite directions, r(48) = -.90*** [-.94, -.83], showing that Foils derived from harder (less-frequently claimed) Reals were more likely to be claimed. There was also an inverse relationship for performance (IT.A), r(48) = -.38** [-.59, -.11], indicating that if one of the pair performed well, the other didn't. Altogether this suggests

that generating Foils to lexically match Reals in the same test set was not a productive approach, because increasing the value of some would decrease the value of others. It also suggested that Reals and Foils need to be optimized independently. These patterns were similar for the Good subset, although weaker because there are fewer items.

2.3.2 Reliabilities (internal consistency)

Typical reliability assessments (e.g., Cronbach's alpha) tap the mean inter-correlation among items (bumped up by Spearman-Brown). In the signal detection framework, however, an Accuracy score indexes hit rate H relative to false-alarm rate. In multi-category versions of overclaiming questionnaires, different domains (e.g., science, arts, music) can serve as statistical units (Paulhus, 2011). In the VOCE, of course, there is only one domain, so the entire survey is effectively a single item.

Instead, we simply reversed the Foils and treated all items as comparable statistical units in computing Accuracy. For Bias, no reversals are necessary. Taking this approach, Cronbach's alpha for the Study 1 item set = .85. With all items scored as positive, they contribute to Bias, with alpha = .84.

Considering only the Good item set, alpha for Accuracy = .61, and for Bias = .76. Using the Spearman-Brown prophecy formula to adjust for number of items, Good items alpha for Accuracy = .80, and for Bias = .89.

2.3.3 External Analyses

In Table 1, a number of demographic variables were able to serve as external validation criteria. Once again demonstrating the impact of ceiling and floor effects, correlations with other measures from the Good subset were always equal to or greater than those from the complete set.

Thus, for brevity, only correlations between other measures and the 20 items of the Study 1 Good set are shown in Table 1.

Two associations were anticipated: (1) VOCE Accuracy should increase with age and (2) VOCE Accuracy should increase with acculturation to Canada. No predictions were made regarding VOCE Bias scores.

2.3.3.1 Demographic Patterns

Table 1 lists demographic variables with the number of valid records, the mean and standard deviation, and correlations with overall VOCE scores. Other than Age, all demographics measured in this study were dichotomous: Hence the mean is the proportion. Our sample had only 11 Japanese (0.38%). Some respondents chose not to categorize their gender. Age was positively skewed, as is typical in student samples.

To preview the results, Accuracy tracked age, as expected. In addition, Accuracy was also associated with variables tapping acculturation, as expected. No differences were found for gender.

	N	*M* (*SD*)	H	F	A	B
Age	2842	20.2 (2.80)	.06**	-.07***	.10***	-.00
Female	2901	.71 (.46)	-.02	.01	-.02	-.01
Native English (vs. ESL)	2922	.74 (.44)	.29***	.02	.25***	.21***
Caucasian	2922	.33 (.47)	.20***	.01	.17***	.15***
English Caucasian	2922	.31 (.46)	.20***	.02	.16***	.15***
Asian	2922	.51 (.50)	-.20***	-.04*	-.15***	-.16***
English Asian	2922	.31 (.46)	.03	-.02	.04*	.01
Japanese	2922	.00 (.06)	-.03	.00	-.03	-.02

Table 1. Study 1 correlations between other measures and VOCE scores.

For dichotomous variables (all except Age), the mean represents proportion.

***Note.* *$p < .05$, **$p < .01$, ***$p < .001$.**

2.3.3.2 Associations with Culture

Being a native English speaker or Caucasian (95% of Caucasians in our sample were native English) positively predicted both Accuracy and Bias. Although being Asian predicted in the opposite direction, it is notable that this effect disappears (and even reverses) for native English speaking Asians (61% of Asians in our sample): In short, language ability carries more weight than culture for VOCE measures. Note similar null effects for Japanese nationality.

Age was correlated with being Asian, $r(2840) = -.12$*** [-.16, -.09]. However, in a regression model predicting Foil claiming from Age, whether a native English speaker and whether Asian, only Age remained significant.

To illustrate this pattern more clearly, Figure 8 compares relative rates of claiming of Reals (H) and Foils (F) based on language ability (native vs. non-native English), culture (Caucasian vs. Asian), and the combination of the two (because Bias considers indiscriminate claiming of either Reals or Foils, simply comparing A and B would obscure these findings). The interesting pattern is that, whereas H varies by language and culture (indicative of the corresponding differences in Accuracy), Foil claiming seems unaffected by either language ability or culture. This can be illustrated by comparing mean Foil-claiming for what should be the two most disparate sub-groups, native English-speaking Caucasians and English as a second language (ESL) Asians, where the difference is .00 [-.01, .01], $t(1176.40) = 0.17$; $d = 0.01$. This pattern is the same for the Good item subset.

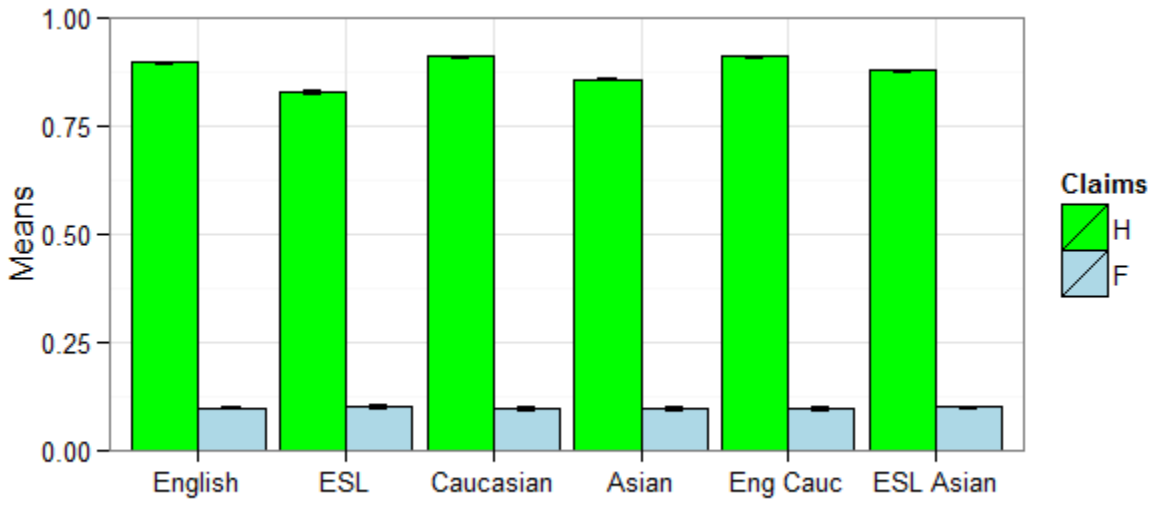


Figure 8. Study 1 Cultural invariance of Foil claiming.

A comparison of H (Reals claiming) and F (Foils claiming) for groupings based on language ability (English / ESL), culture (Caucasian / Asian) and both (Eng Cauc / ESL Asian). Error bars show standard error (which are very small, since N = 2922).

2.3.3.3 Predictors of Item Behavior

A key aspect of this research program was to identify what characteristics make for better overclaiming items. Hence, in this section, item claim rates and performance were correlated with item properties to search for patterns. The analyses begin with all 25 Reals and 25 Foils.

2.3.3.3.1 Reals

As expected, Real claim rates correlated substantially with the Difficulty of the word, as rated by the Top 1000 list, $r(23) = -.85$*** [-.93, -.69]. However, this pattern did not hold for the Good items, $r(8) = -.08$ [-.68, .58], which may be due to range restriction, since all 10 Good Reals had Difficulty in the top 100 (i.e. were "Hard").

Noting that the mean difference in claim rates between Hard ($M = .71$, $SD = .14$) and Easy ($M = .99$, $SD = .00$) Reals was .28*** [.18, .38], $t(9.01) = 6.32$; $d = 2.87$. Note that the 15 Easy Reals had virtually identical claim rates, making them of little value for discriminating ability.

The ELP measures provided no further clarification. The one exception was a trend indicating that how frequently a word appears in print might be useful in refining word difficulty.

2.3.3.3.2 Foils

For Foils, length of the item was found to influence performance, $r(23) = .40$* [.01, .69], suggesting that longer Foils have more discriminability, although length had no impact on claim rate, $r(23) = .03$ [-.37, .42]. Unfortunately, no other significant lexical predictors were found for Foils in this set, although the power to test this was limited by having only 25 items, and the fact that Foils were selected based on lexical properties.

2.4 Discussion

Our initial foray into systematically choosing items for vocabulary overclaiming proved fruitful, illuminating both dead ends and possible new paths.

The notion of generating Foils to match Reals apparently was a disappointment: However it did suggest that Foils that resemble a more challenging genuine word were more likely to be claimed. Reals clearly need to be more challenging to gain useful variance. Longer Foils may perform better.

Naively choosing items for college samples can easily result in ceiling effects for Reals: The present results show how those problems can distort predictions, highlighting the need for research like this. More encouraging was the finding that significant demographic relationships could be found using only 20 items, suggesting that, with well-chosen items, overclaiming measures could indeed be made more efficient. The challenge now was to refine the process of choosing optimal items.

Chapter 3: Study 2

A central lesson from Study 1 was to avoid ceiling and floor effects. Floor effects for Foils may be an eternal challenge, since claiming familiarity with non-existent items will always be relatively rare. Study 2 will focus on increasing the number of VOCE items while matching the quality of the so-called 'Good' set of 20 items Study 1.

3.1 Item Selection

3.1.1 Reals

Because Reals from the lower 100 of the Top 1000 list were too easy, we chose Reals from the 600-900 range – but with twice as many in the 900-1000 range, since this is where the Study 1 Good set came from. We then used the word frequency measure from English Lexicon Project (ELP) to fine tune the selection. These Reals were expected to be less constrained by ceiling effects and thus generate more variance, in hopes of improving overall performance.

3.1.2 Foils

Using Wuggy to model Foil qualities after our Real qualities was of limited value. Instead, we turned to the ELP because it provided a large set of non-words. From that archive, we chose words that were not too short and with 3 or more "orthographic neighbors": That is, nonwords having genuine words that differ by one letter. Such Foils should be more seductive.

Unfortunately, a large portion of such ELP nonwords had very common endings, such as "ing", "er", "ed", or those with an additional "s". From 255 results, only 6 had other endings. To avoid cueing Foils by their suffix, I selected a diversity of suffixes.

3.2 Method

3.2.1 Participants and Procedure

The same method of gathering participants was used as for Study 1. The final sample size was 3242. Mean age was 20.1 with 72 percent female respondents. More details about the participant sample are provided in Table 2.

3.2.2 Measures

The VOCE part of the prescreen was again 50 items: 25 Reals and 25 Foils, in random order. This time, however, the stem was changed to encourage interest: "Typical students at this university have exceptional vocabularies. Rate your familiarity with the English words below". We also changed the claiming option from "Heard of it somewhere" to the more definitive "Have heard of it". An interim pilot test had showed that changing the stem had no effect on overall Bias (.00 [-.00, .01], $t(1297.05) = 0.83$; $d = 0.03$) but a slight drop in Accuracy (-.02*** [-.04, -.01], $t(1241.28) = -3.50$; $d = 0.15$).

Demographic items in the prescreen were similar to those in Study 1 Unfortunately, no language background information was included.

3.3 Results

Again, only respondents with full VOCE data were used. While only 1.97% of subjects had negative Accuracy (compared to 0.21% in Study 1, probably because those Reals were so easy), one subject (the only Male Caucasian with negative Accuracy) claimed no Reals and 87.5% of the Foils, so his results were removed as an outlier, leaving 3242 valid survey responses.

3.3.1 Internal Analyses

3.3.1.1 Distributions of H and F

Study 2 results no longer showed ceiling effects for H (M = .70, SD = .21), with mean Reals claiming falling significantly compared to Study 1, by -.17*** [-.18, -.17], t(4872.62) = -41.10; d = 1.02 (see Figure 9). Foil claiming was still compromised (M = .15, SD = .18) although it was higher than Study 1, by .05*** [.04, .06], t(5993.40) = 12.72; d = 0.32. 24% of respondents claimed no Foils at all.

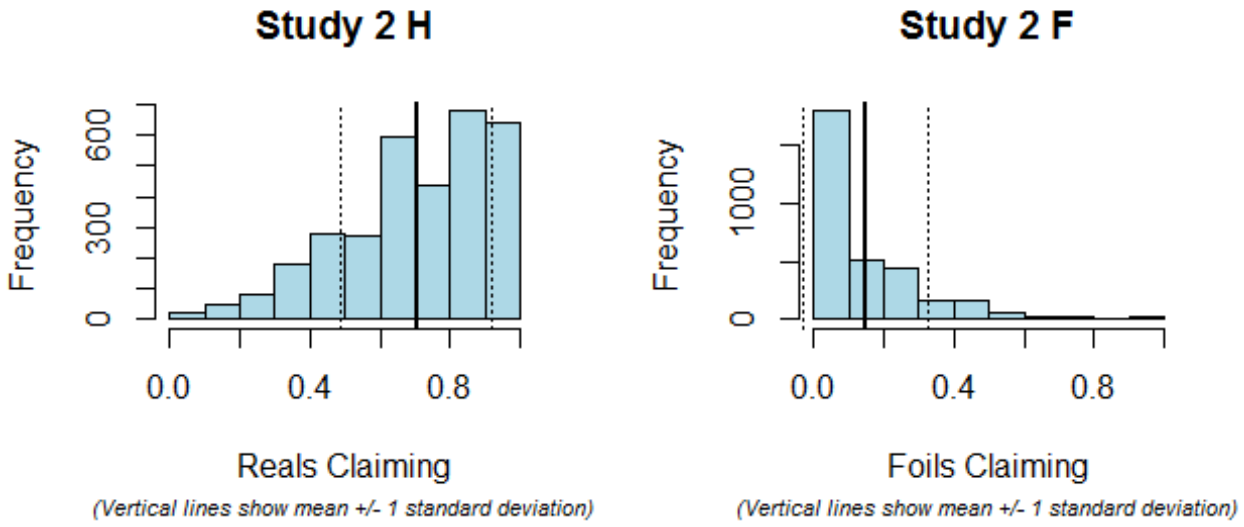


Figure 9. Study 2 distributions of per-subject claiming of Reals and Foils.

Figure 10 shows claim rates for individual items shows the same pattern: Reals claiming had a wider spread, while Foil claiming improved slightly over Study 1. Interestingly, the Real with the highest claim rate (and lowest performance), "advocated", was originally the foil "alvocated", but was mistakenly "corrected" in assembling the survey.

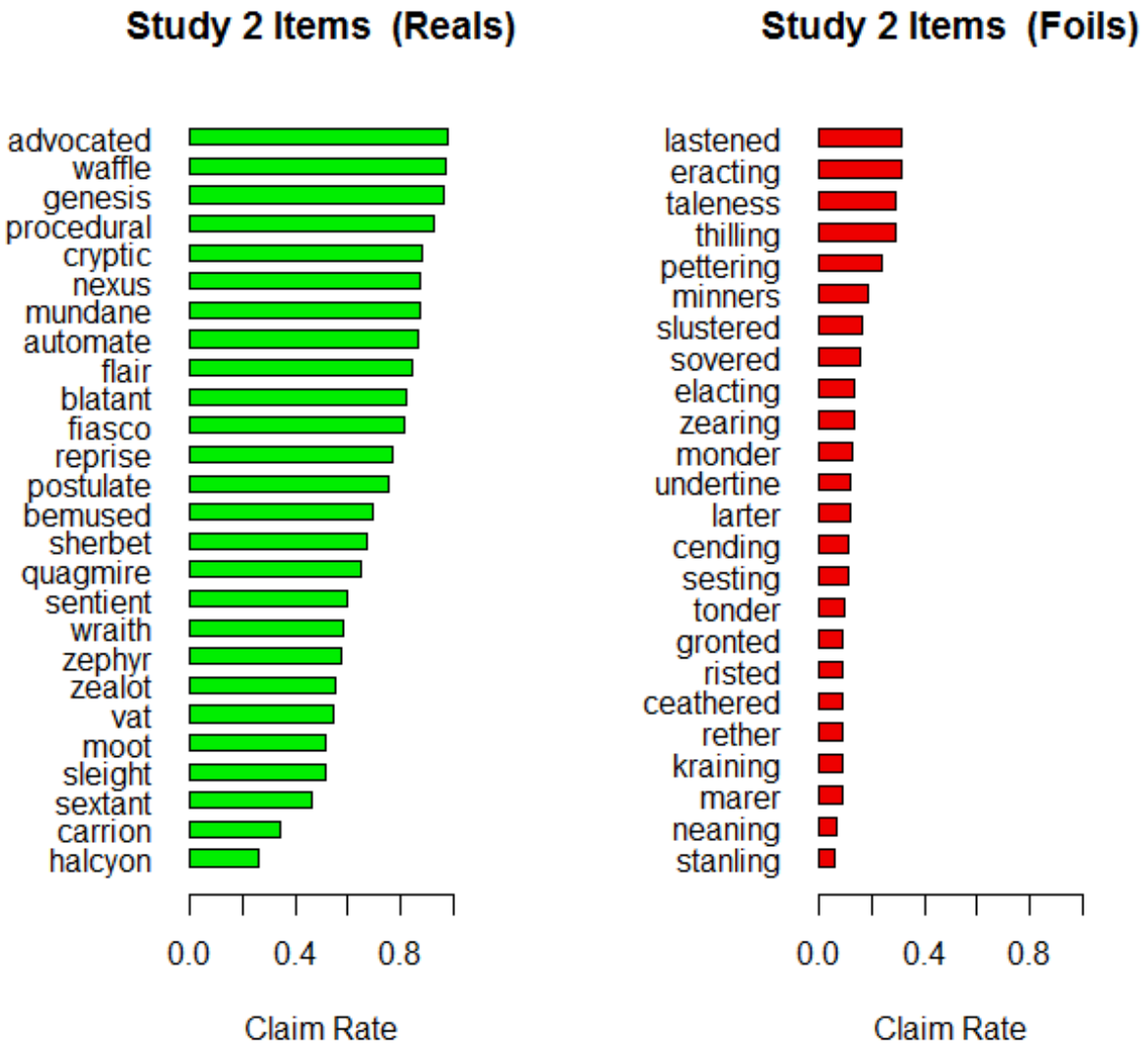


Figure 10. Study 2 Real and Foil items ranked by claim rates.

3.3.1.2 Distributions of A and B

Bias (M = .43, SD = .15) was lower than Study 1 by -.06*** [-.07, -.06], t(5473.81) = -19.81; d = 0.49, as was Accuracy (M = .55, SD = .26), by -.23*** [-.24, -.21], t(5503.31) = -41.61; d = 1.04, with Accuracy showing a broader and less skewed distribution (see Figure 11).

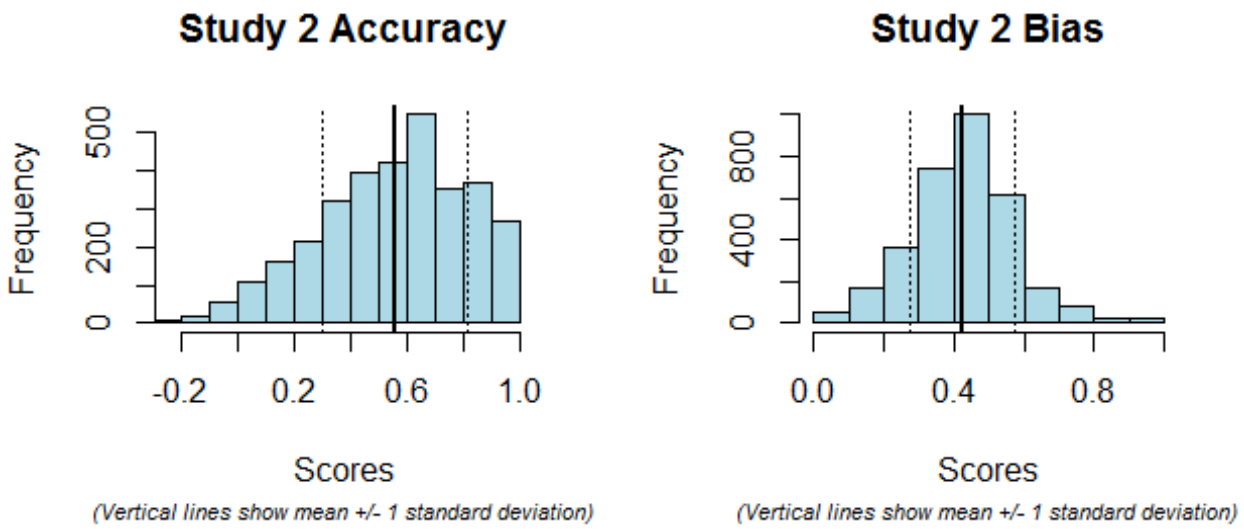


Figure 11. Study 2 distributions of per-subject Accuracy and Bias.

3.3.1.3 Correlations

H and F correlate positively ($r(3240) = .14$*** [.11, .18]), similarly to Study 1. Unlike Study 1, A and B correlated positively ($r(3240) = .18$*** [.15, .21]) without removing items, suggesting that at least this aspect of ceiling effects had been corrected.

3.3.1.4 Performance

As seen in the Study 1 results, claim rates again did not predict performance for either Reals, $r(24) = .01$ [-.38, .40], or Foils, $r(22) = .04$ [-.37, .43].

Unlike the Study 1 entire item set, but like the Study 1 Good subset, performance was lower for Foils ($M = .30$, $SD = .06$) than for Reals ($M = .38$, $SD = .09$), with a difference of .08*** [.04, .13], $t(43.93) = 3.80$; $d = 1.07$.

In Figure 12, we can also see a wider spread of claiming and performance, especially for Reals. Compared to Study 1, there was an increase in performance for Reals of .12*** [.07, .16], $t(41.74) = 5.47$; $d = 1.53$ but a decrease for Foils of -.07*** [-.10, -.03], $t(46.67) = -3.87$; $d = 1.11$.

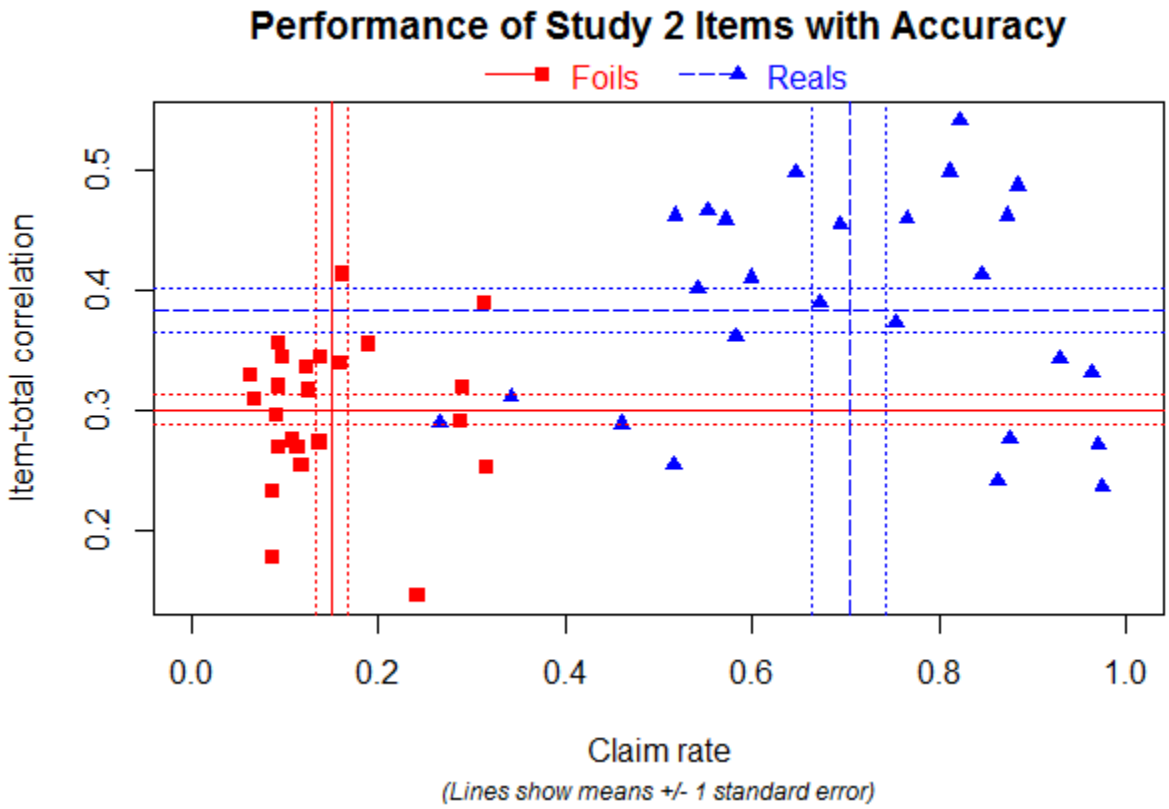


Figure 12. Study 2 Claim rates vs Performance.

3.3.1.5 Reliabilities

Using the same technique as we did for Study 1, the internal consistency value for Accuracy was .85; for Bias it was .89.

3.3.2 External Analyses

3.3.2.1 Demographic Patterns

The Study 2 data set did not include items for language background, so analysis of culture invariance was not possible. As before, all measures other than Age are dichotomous, so means represent proportions. See Table 2.

	N	*M* (*SD*)	H	F	A	B
Age	3167	20.10 (2.86)	.08***	-.07***	.12***	.02
Female	3222	.72 (.45)	-.17***	-.02	-.13***	-.13***
Caucasian	3242	.30 (.46)	.25***	.03	.18***	.20***
Asian	3242	.52 (.50)	-.21***	-.06***	-.13***	-.19***
Caucasian Female	3222	.22 (.41)	.15***	.02	.11***	.12***

Table 2. Study 2 correlations between demographics and VOCE scores. For dichotomous variables (all except Age), the mean represents proportion.

***Note.* *$p < .05$, **$p < .01$, ***$p < .001$.**

Again, Age tracked Accuracy. The group of Caucasian Females was included to show that gender was confounded with ethnicity, which was likely also confounded with language ability, which can't be isolated in this data set.

3.3.2.2 Predictors of Item Behavior

3.3.2.2.1 Reals

As expected from Study 1 results, Reals claim rates no longer correlated significantly with the Difficulty of the word, $r(23) = -.19$ [-.55, .22], since they were now restricted to the 600-1000 range (vs. the full range in Spring). This also indicates that Difficulty (as measured by rank in the Top 1000 list) was a crude measure of how challenging a word is for a vocabulary test.

3.3.2.2.2 Foils

For Foils, length of the item no longer had an impact on performance, $r(22) = -.00$ [-.40, .40], although it marginally affected claim rate, $r(22) = .40$ [-.00, .69].

Overall psycholinguistic predictors were not conclusive. Patterns found over multiple studies are discussed in a later chapter.

3.3.3 Comparison Between Studies

Compared to Study 1, performance of Reals was significantly improved (by .12*** [.07, .16], $t(41.74) = 5.47$; $d = 1.53$), probably because ceiling effect was substantially mitigated. The mean claim rate of all Reals in the Study 2 set was essentially equivalent to that of the Study 1 Good set -.00 [-.02, .01], $t(5940.20) = -0.84$; $d = 0.02$.

Foil claim rates were also improved, almost to the levels of the Study 1 Good set (a difference of -.01* [-.02, -.00], $t(5902.18) = -2.58$; $d = 0.07$). Reliabilities remained high.

3.4 Discussion

Now that the entire item set was behaving like the Study 1 Good subset, we could reasonably assume that the overall instrument was strong enough to be tested as an overclaiming test. We were also interested in how well Reals claiming could be controlled and if any reliable predictors of Foil behavior could be found. Furthermore, the finding of cultural invariance of Foil claiming needed to be replicated.

Chapter 4: Study 3

Having explored item selection techniques in Study 1 and Study 2, and developed what appeared to be a good overclaiming instrument, dubbed the VOCE. The next step was to confirm the patterns we found and validate the measure.

4.1 Goals: Confirm and Validate

We had several goals for Study 3. With a better idea of what kind of Reals to choose, we should be able to fine tune the distribution of H (Reals claiming). Without conclusive indicators for Foils we could keep exploring. Nonetheless, our item set seemed good enough to warrant evaluating the measure overall.

To that end, we designed a study to 1) retest Study 2 items in a different context (test-retest reliability), 2) compare that item set with another set generated by similar techniques (concurrent reliability), and 3) evaluate convergent validity of Accuracy and Bias with other relevant measures.

4.2 Item Selection

For retest and concurrent reliability, we kept the Study 2 items (with identical prompt and choice options) and created a parallel set.

4.2.1 Reals

The "New" Reals (as opposed to "Repeat" items) were chosen in a similar fashion to Study 2, but with a slight tweak to test our control in item selection. Overall Reals claiming (H) in Study 2 was .70 (.21) which still seemed a bit high, so we aimed to lower that by choosing items with Difficulty only in the 900s range (vs. 600-900 in Study 2).

The one exception to exact replication of Study 2 items was the misprint "advocated" which was intended to be the Foil "alvocated", thus making Study 2 inadvertently contain 26 Reals and

24 Foils. Rather than retain the unintended (and poorly performing) “advocated”, we used “alvocated”: Hence the New set now has 26 Foils and the Repeat set has 24 Foils. Both sets had 25 Reals, so New set had 51 items while the Repeat set had 49 and “advocated” was not repeated.

4.2.2 Foils

Given the apparent restrictions on Foils found in the ELP database, we tried another software application designed to generate nonwords for psycholinguistic research, namely, WordGen. Unlike Wuggy, words can be generated given lexical properties alone – without using genuine words as templates. We chose words with length of 7-10 letters (because length was positively associated with performance in Study 1) and with 4 or more orthographic neighbors, because we still believed that similarity to English words might help.

4.3 Method

4.3.1 Participants & Procedure

We developed an online survey using Qualtrics software (Qualtrics, Provo, UT) to facilitate flexible yet consistent delivery. This allowed us to run participants in a laboratory setting, or let them take the survey online. The population was still undergraduate psychology students, but now, rather than taking a prescreen survey as part of joining the HSP, they had to actively choose to participate in our study, described as being about vocabulary and personality.

The final sample size was 338 undergraduates. Mean age was 20.4 and percent female was 66. More details about the sample are provided in Table 5.

Our study had the same course credit incentive as the prescreen, but we now controlled all the content directly on Qualtrics. Students could choose either an online version or an in-lab version. Both involved the same survey; only the context varied.

Although demographic questions always appeared at the start of the survey, the order of other components was counterbalanced. VOCE item order was randomized for each participant.

4.3.2 Measures

To allow us to match respondents' new data with their prescreen data, participants were asked to provide their HSP IDs and were reminded (with identical instructions from the prescreen) how to construct their IDs.

4.3.2.1 Demographics, Culture and Language

To make comparisons with the cultural findings in Study 1, we included self-reported ethnicity and the number of years they'd spent in Western countries. Also included were age, gender, and number of completed University years.

To assess proficiency with English, we asked respondents to report the age they were first comfortable speaking English, how *experienced* they felt reading and writing English, and how *confident* they felt reading and writing English.

4.3.2.2 Vocabulary Ability Measure

To assess language ability with a standard approach, we included an independent vocabulary test. The UBC Word test is a 50-item multiple-choice quiz asking the subject to select an appropriate synonym from four options. It was modeled after the Quick Word test developed by Borgatta & Corsini, (1960) and correlates .55 with the Wonderlic IQ test, and .67 with its Verbal subscale (Nathanson & Paulhus, 2007). Mean score of the 100-item version was 35.3 with a standard deviation of 12.9 in a large UBC undergraduate sample. The alpha was .86.

4.3.2.3 Personality Measures

To see how VOCE scores related to personality factors, we included a variety of relevant personality measures. The Balanced Inventory of Desirable Responding (BIDR) was included as

a self-enhancement measure because it has separate measures (20 items each) for both impression management (IM) and self-deceptive enhancement (SDE) (Paulhus, 1984). The 40-item Narcissistic Personality Inventory (NPI; Raskin & Hall, 1979) and the 16 items from the HEXACO Honesty-Humility (HH) scale (Ashton & Lee, 2009) were also included.

4.3.2.4 Follow-up Items

At the end of the survey were included two "follow-up" questions to assess subject perception of the VOCE items: "What percentage do you think were NOT genuine English words?", and "What percentage had you seen before in another study with the same question?" The latter is relevant because anyone who had done the prescreen would have seen half the items before (the Repeat set).

4.3.3 Hypotheses

We anticipated several results.

4.3.3.1 H1: Better distributions H and F

We hoped for better overall claiming of Foils (F), although the basis for this was more intuitive than empirical. Based on the above strategy, we did have good reason to predict that Real claim rates (H) should be similar to those in Study 2 for the Repeat set and lower for the New set, making for a broader, more centralized (closer to .5) distribution of H overall.

4.3.3.2 H2: Convergent Validity - Capturing Language Ability

Given that VOCE's content domain is vocabulary, Accuracy scores should closely track performance on the UBC Word test. There should also be strong correlations with the other language ability measures described above (e.g., English as native language).

4.3.3.3 H3: Cultural Invariance - Replication of Study 1 findings

If the pattern found in Study 1 was valid, Foil claiming would show no correlation with ethnicity or language ability, whereas Real claiming would.

4.3.3.4 H4: Personality Correlates

Overclaiming is presumed to tap self-enhancement, and previous research has shown correlations with narcissism and social desirability measures (Paulhus et al., 2003). We predicted that VOCE would do likewise. We were agnostic about links with Honesty-Humility but it seemed plausible that it should be inversely related to Foil claiming.

4.3.3.5 H5: Accountability

Studies 1 and 2 were done via HSP prescreen which is arguably the most accessible "study" on the system since it is directly advertised, available immediately and requires no sign up, scheduling or debriefing. Even a casual perusal of prescreen results shows a number of students don't take it very seriously, skipping items or answering insincerely. While problematic results can easily be discarded (given thousands of valid results), it is reasonable to assume that students don't take the prescreen as seriously as a study they actively select and sign up for. Involvement rises further when a student has to make an appointment and come into a lab to take a survey. Presumably, increased involvement increases sincerity and sense of accountability. We chose to test these three levels of accountability. Since we already had prescreen data, Study 3 would collect data on the same items both for online administrations (but where the student signed up and agreed to come later to a debriefing) and in-lab administrations. Based on previous research (Paulhus et al., 2003), we predicted that Bias would decrease in a context of greater accountability.

4.3.3.6 H6: Previous Exposure

Some of our sample will have seen half the items (the Repeat set) before in the prescreen. One might reasonably predict that this previous exposure would increase claiming overall (of the Repeat set), including of Foils (Williams et al., 2002).

4.4 Results

In total, 338 usable surveys were collected. Missing data was not an issue because all items required a response. However, some ethnicity items and two follow-up items weren't reported on some surveys. No outliers were identified.

Reporting of results are grouped into three sections: those about the behavior of VOCE items themselves (Internal Analyses), its internal, test-retest and concurrent reliabilities (Reliabilities) and how VOCE measures compare with other measures (External Analyses).

4.4.1 Internal Analyses

4.4.1.1 Distributions of H and F

Figure 13 shows that the Repeat set distributions of H (M = .72, SD = .23) and F (M = .21,

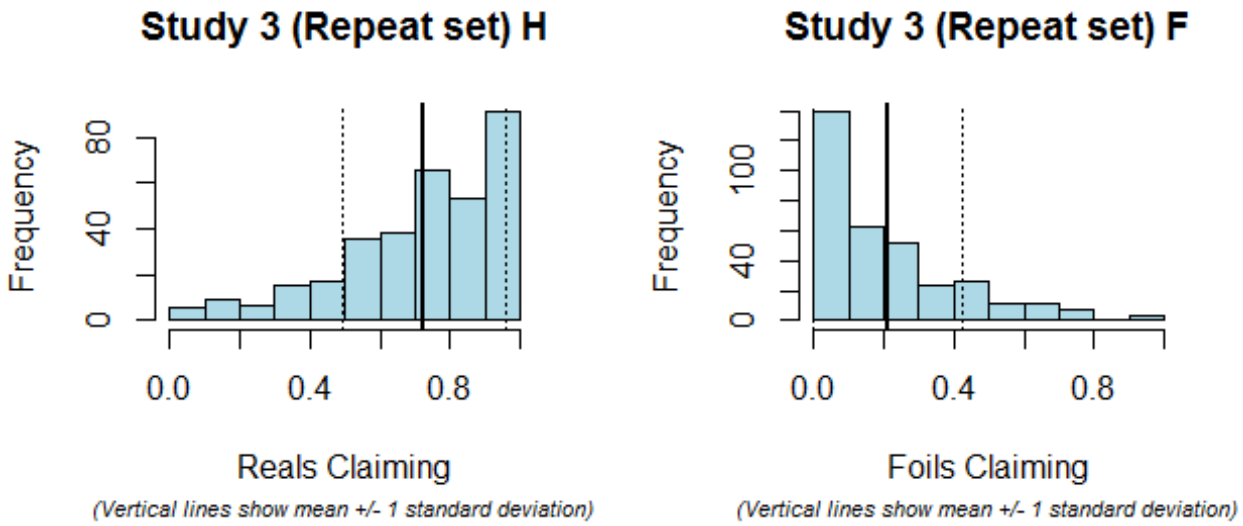


Figure 13. Study 3 Repeat set distributions of per-subject claiming of Reals and Foils.

SD = .21) were, as expected, similar to those of Study 2 (since there were the same items).

The New set distributions of H (M = .43, SD = .21) and F (M = .24, SD = .20) were shifted more centrally (closer to .5), as intended. See Figure 14.

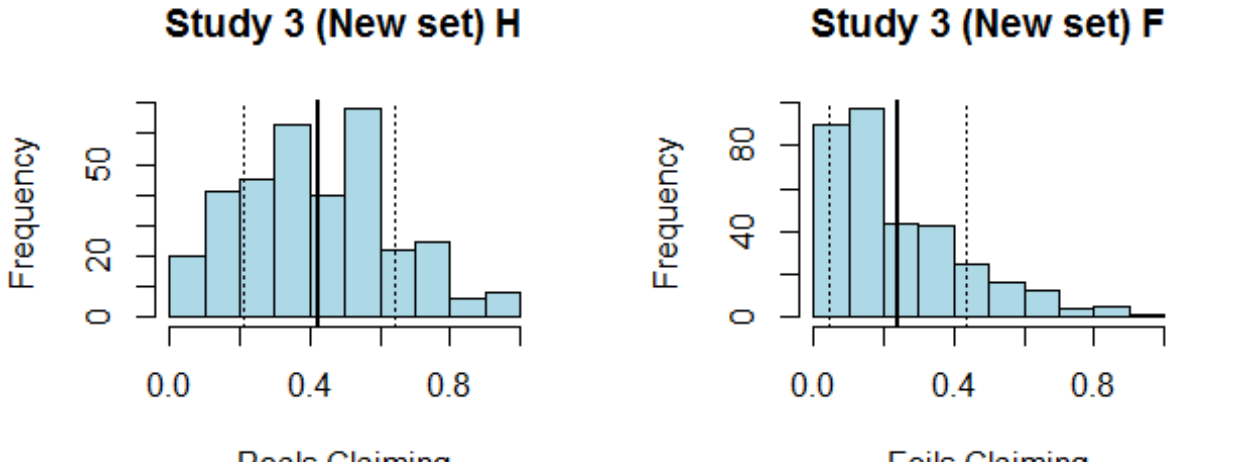


Figure 14. Study 3 New set distributions of per-subject claiming of Reals and Foils.

Together, Study 3 results were not nearly as cramped by ceiling or floor effects for H (M = .57, SD = .21) or F (M = .23, SD = .20) respectively (Figure 15). Reals claiming (H) now shows a good match of item difficulty to sample ability, un-hindered by range. Foils claiming now covers the full range, with a mean more than one standard deviation from the lower limit.

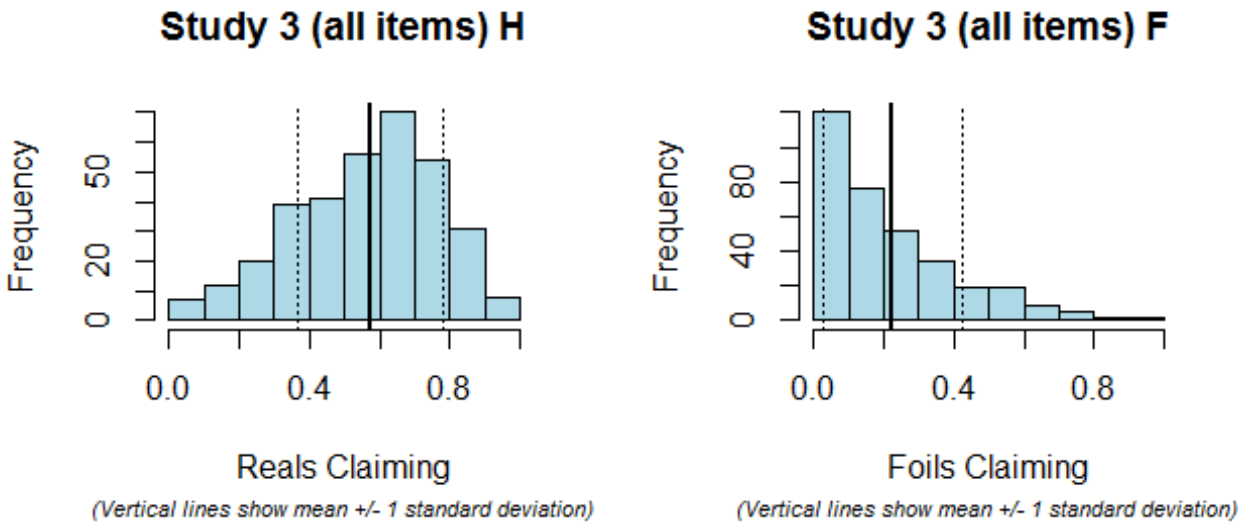
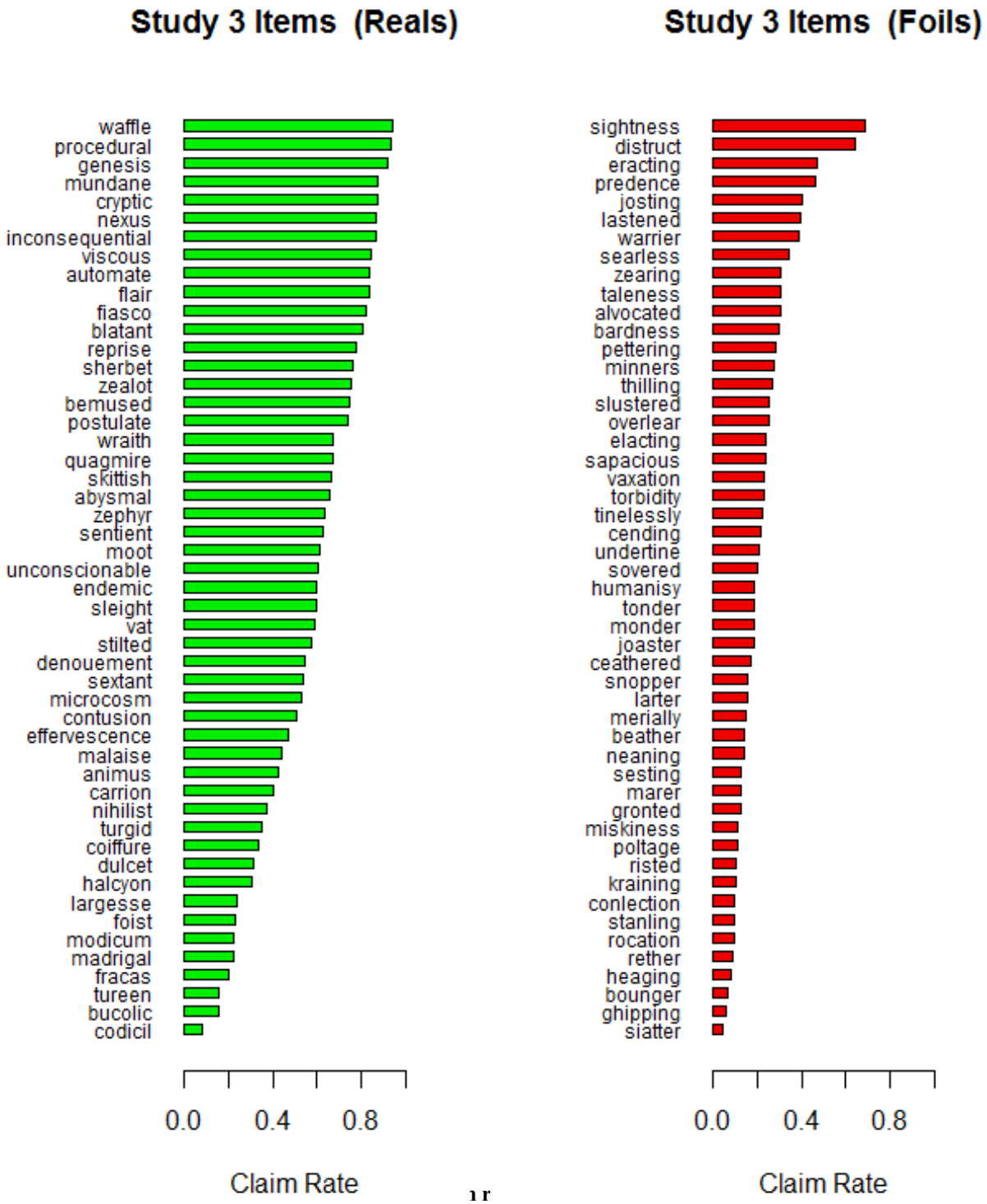


Figure 15. Study 3 distributions of per-subject claiming of Reals and Foils.

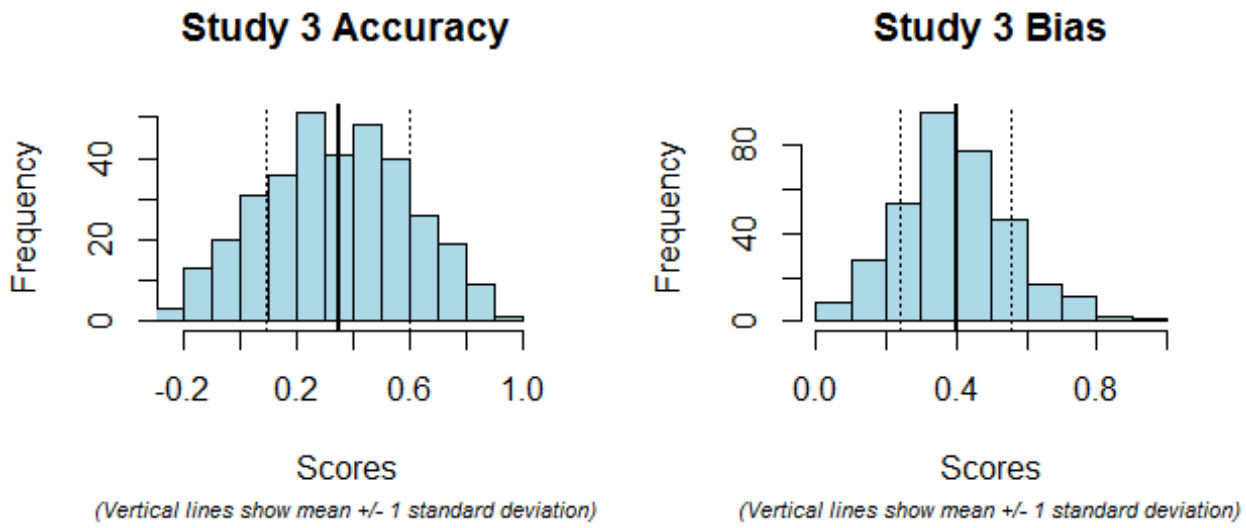
Looking at claim rates for individual items (Figure 16) we now see a much wider and more even spread of difficulty, or in Item Response Theory terms, location along trait spectrum.



Only 3.85% of respondents claimed no Foils at all. Keeping in mind that there were twice as many items, for fair comparison, the no Foil claiming percentage for the Repeat and New items sets were 16% and 7% respectively – still quite an improvement over Studies 1 and 2. Recall that other overclaiming applications have reported 30% (or more) of respondents claiming no Foils, leaving a large portion of the sample insensitive to the instrument.

4.4.1.2 Distributions of A and B

We had sought claim rates closer to 50% (i.e., more central). Indeed, more central distributions (Figure 17) were found for Bias (M = .40, SD = .16), and Accuracy (M = .35, SD = .25). The negative Accuracy tail, assumed to be noise, seems to fit within the distribution, with range similar to Studies 1 and 2, yet without ceiling effects.



4.4.1.3 Correlations

Figure 17. Study 3 distributions of per-subject Accuracy and Bias.

As expected, H and F correlated positively (r(336) = .22*** [.12, .32]). A and B showed no correlation (r(336) = .05 [-.06, .15]).

4.4.1.4 Performance

Claim rates strongly predicted performance for Reals, $r(48) = .59^{***}$ [.38, .75], suggesting harder Reals performed less well, unlike in Study 2. However, this relationship was not significant for the Repeat set, nor did the Repeat set have a significant difference in performance from Study 2. Since the relationship *was* significant for the (intentionally harder) New set ($r(23) = .48^{*}$ [.10, .73]), and it had significantly lower performance (shifting -.11** [-.18, -.03], $t(39.20) = -2.94$; $d = 0.83$), it seems likely that some of the New Reals were too difficult in that they compromised performance.

There was no relationship found between claim rates and performance for Foils, $r(48) = -.02$ [-.29, .26]. Overall, performance of Foils ($M = .31$, $SD = .09$) was equivalent to that of Reals ($M = .31$, $SD = .17$), with a difference of .00 [-.05, .06], $t(75.63) = 0.15$; $d = 0.03$. This seems to be more because of the drop in Reals performance.

Figure 18 shows the dispersion of claiming and performance. Note that there are four Reals with negative performance; they were all from the New set.

4.4.1.4.1 Negatively Performing Items

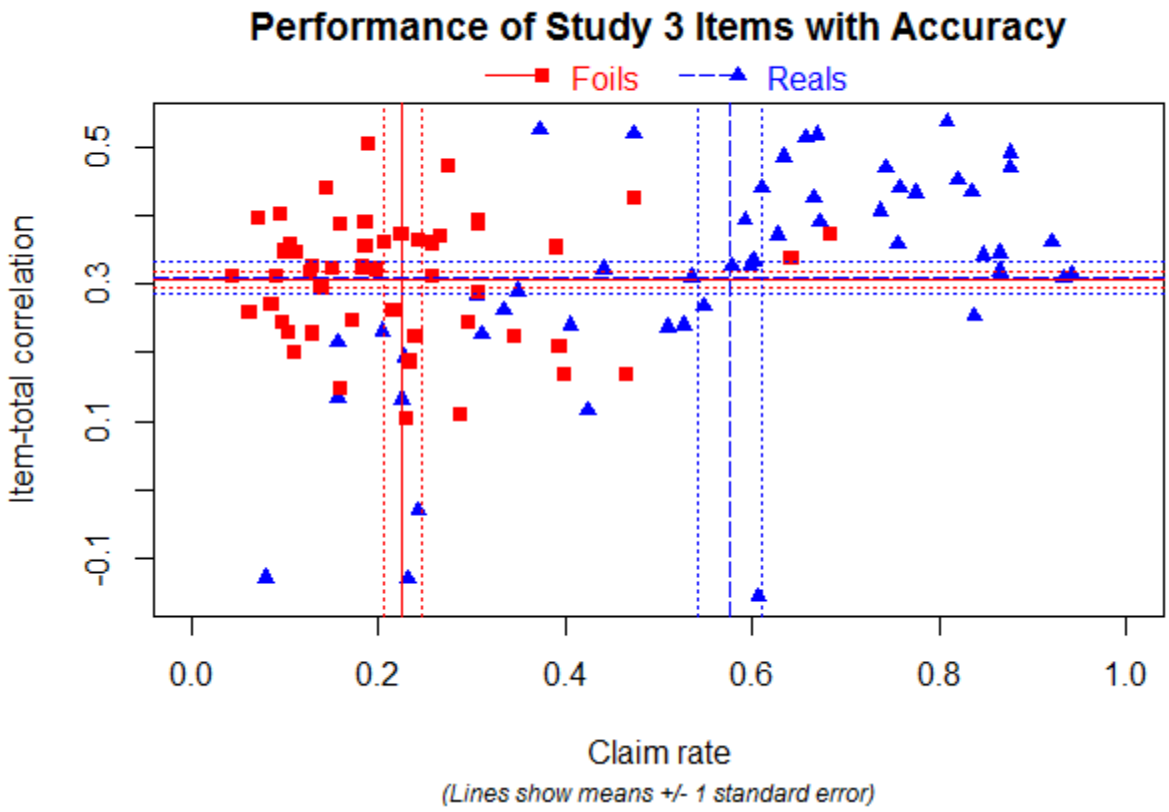


Figure 18. Study 3 Claim rates vs Performance. Note the four New Reals with negative performance.

The only items whose claim rates were negatively related to overall Accuracy were the Reals *unconscionable, foist, codicil, largesse* (all "New" items, worst listed first). Thus they acted like Foils (and having claim rates also correlated most strongly with overall Foil claiming). The last three have low claim rates and weak lexical predictors; they are uncommonly difficult words, probably too challenging for this sample. However, "unconscionable" was correctly identified as Real by 61%. Despite that one anomaly, these results do suggest the importance (and feasibility) of carefully matching familiarity of Reals to the ability of the sample.

4.4.2 Reliabilities

This section reports how VOCE evaluates two forms of reliability: test-retest of same items at different times or contexts, and parallel tests with different items.

4.4.2.1 Internal

Using the same approach as for Studies 1 and 2, Cronbach's alpha for Accuracy in Study 3 item set = .91, and for Bias = .94. For the Repeat set alone, alphas were .87 and .89 for Accuracy and Bias, respectively. For the New set, .77 and .90.

With the negatively performing Reals removed, New set reliabilities for Accuracy and Bias become .81 and .89, respectively. Subsequent analyses are after those four New Reals have been removed.

4.4.2.2 Test-Retest

Out of 338 observations, we found 262 (78%) that matched appropriate HSP IDs with the Study 3 survey, which allowed us to test the (presumably) same subject on the same items at a later time. Correlations and differences are shown in Table 3.

However, since HSP IDs are not necessarily unique, correlating responses based on this will certainly include some error, which would dampen test-retest correlations in the table below. (See below for more possible HSP ID matching confounds.) The means differences below are calculated using a paired t-test. The increase in claiming may be due to change of context; see the results on Accountability, below.

Measure	Correlation	Change from Study 2 to 3
H	$r(260) = .77$*** [.72, .82]	.02* [.01, .04], $t(261.00) = 2.58$; $d = 0.11$
F	$r(260) = .50$*** [.40, .58]	.04*** [.02, .07], $t(261.00) = 3.85$; $d = 0.24$
A	$r(260) = .70$*** [.64, .76]	-.02 [-.05, .00], $t(261.00) = -1.63$; $d = 0.08$
B	$r(260) = .61$*** [.53, .68]	.03*** [.02, .05], $t(261.00) = 4.19$; $d = 0.23$

Table 3. Comparing VOCE scores by subject between Studies 2 and 3.

Note. *p < .05, **p < .01, ***p < .001.

Ignoring subject matching, per-item claim rates showed remarkable consistency between the two administrations, $r(47) = .99$*** [.98, .99], while item performance varied only slightly, $r(47) = .79$*** [.65, .88].

4.4.2.3 Concurrent

Typically concurrent reliabilities are based on correlations between parallel tests. Here we're using the term "parallel" to indicate two test versions that are highly inter-correlated with similar variances but with different means. Thus, we're testing these two item sets (Repeat and New) as parallel, even though the latter was intended to have lower means from more challenging items.

The above correlations are based on the same items but tested at different times. Compare that with Table 4 correlations between testing subjects at the same time but different items; the Repeat set vs the New set. It appears that our item selection technique introduces less inconsistency than time and/or context does. The noticeable drop in Reals claiming (and thus Accuracy) is likely due to our choosing more difficult Reals. Paired t-tests are used for differences.

Measure	Correlation	Difference from Repeat (old) to New
H	$r(336) = .95$*** [.94, .96]	-.13*** [-.13, -.12], $t(337.00) = -31.53$; $d = 0.56$
F	$r(336) = .97$*** [.96, .98]	.01*** [.01, .02], $t(337.00) = 5.25$; $d = 0.07$
A	$r(336) = .96$*** [.95, .97]	-.14*** [-.15, -.13], $t(337.00) = -29.45$; $d = 0.49$
B	$r(336) = .96$*** [.95, .97]	-.06*** [-.06, -.05], $t(337.00) = -22.13$; $d = 0.34$

Table 4. Comparing VOCE scores by subject between Repeat and New item sets within Study 3.

Note. *p* < .05, **p < .01, ***p < .001.

4.4.3 External Analyses

Table 5 summarizes the data gathered from other measures, their relationships with VOCE scores, and (for composite measures) their alpha reliability.

	N	*M* (*SD*)	H	F	A	B	α
Age	335	20.37 (3.11)	-.04	-.12*	.06	-.10	
Female	338	.66 (.48)	-.19***	-.06	-.12*	-.17**	
University Year	338	2.66 (1.44)	.01	-.12*	.10	-.07	
Self-report GPA %	338	76.94 (9.83)	.17**	-.06	.18***	.08	
Western (vs Eastern)	198	.39 (.49)	.23**	-.01	.20**	.16*	
Years in Western Countries	335	14.45 (7.72)	.36***	-.11	.37***	.18***	
Age Learned English	338	3.22 (2.82)	-.47***	.04	-.41***	-.30***	
English Experience	338	.92 (.15)	.52***	-.12*	.51***	.28***	
English Confidence	338	.88 (.17)	.54***	-.10	.51***	.30***	
Confidence-Experience	338	-.03 (.08)	.16**	.01	.12*	.11*	
Portion thought fake	112	.43 (.22)	-.12	-.25**	.07	-.22*	
Portion thought seen before	112	.32 (.29)	.06	.18	-.07	.14	
UBC Word Test	338	.60 (.16)	.60***	-.21***	.64***	.28***	.86
Honesty-Humility	338	.56 (.13)	.11	-.12	.18**	.01	.78
Self-Deceptive Enhancement	338	.49 (.09)	-.02	.10	-.09	.05	.67
Impression Management	338	.47 (.12)	-.09	-.03	-.05	-.08	.77
Narcissism	338	.36 (.17)	-.12*	.15**	-.21***	.01	.85

Table 5. Study 3 correlations between other measures and VOCE scores.

For dichotomous variables (where means are < 1), the mean represents proportion.

***Note.* *$p < .05$, **$p < .01$, ***$p < .001$.**

4.4.3.1 Age, University Year

Age was quite positively skewed, as is typical of student samples. While Accuracy was not significantly predicted, Foil claiming was negatively related as expected. Perhaps increased difficulty dampened the Accuracy relationship.

4.4.3.2 GPA (Self-Report)

Previous research showing that that Accuracy tracks cognitive ability is confirmed by the positive correlation with GPA, which logically follows Word score, $r(336) = .20***$ [.09, .30]. While it seems reasonable that GPA positively relates to English experience ($r(336) = .16**$ [.05, .26]) and confidence ($r(336) = .17**$ [.07, .27]), it was unexpected to find a negative relationship with Age ($r(333) = -.28***$ [-.38, -.18]) and university year ($r(336) = -.51***$ [-.58, -.43]), neither of which can be explained by outliers. Also curious is that GPA predicts Honesty-Humility, $r(222) = .26***$ [.14, .38], and declared ethnicity (Western vs Eastern, $r(196) = .14*$ [.00, .28], but not years in Western countries, $r(333) = .02$ [-.09, .13].

Self-report confounds may apply. Since 45% of respondents gave numbers in multiples of 5 (compared to a baseline of 10%, since range was 50-100), we can assume students are estimating their GPA levels. And since reporting a multiple of 5 correlated negatively with GPA ($r(336) = -.18***$ [-.28, -.08]), it appears that higher GPAs were reported more precisely (the three most commonly reported values were 70, 80, and 75, respectively). See Appendix B for further discussion of response rounding.

4.4.3.3 Gender

The poorer performance by females seems surprising, although it appears to be more from lack of claiming overall. In our sample, being female follows University year ($r(336) = -.11*$ [-.22, -.01]), but does not correlate significantly with age, ethnicity, experience, confidence, years

in Western countries or age of learning English. When either H or B are regressed on all those variables, gender remains the most significant predictor, so confounds are not evident.

4.4.3.4 Cultural Invariance of Foil Claiming

Declared ethnicity was 35% Western (European or North American), 53% Eastern (e.g. China, Japan), and 12% Other; our cultural binary measure of Western vs Asian considers only the first two categories. For all respondents, 77% were comfortable speaking English by age 4 (which we consider as native speakers; others are English as a Second Language, or ESL).

In Table 5 above, the one VOCE measure that didn't correlate with cultural or language background was Foil claiming, yet Foil claiming did predict Narcissism. This is the same kind of invariance for F we saw in previous studies.

As in Study 1, comparing the most disparate groups in terms of language and culture (Western native English speakers and Eastern ESL respondents; Figure 19), we see the latter

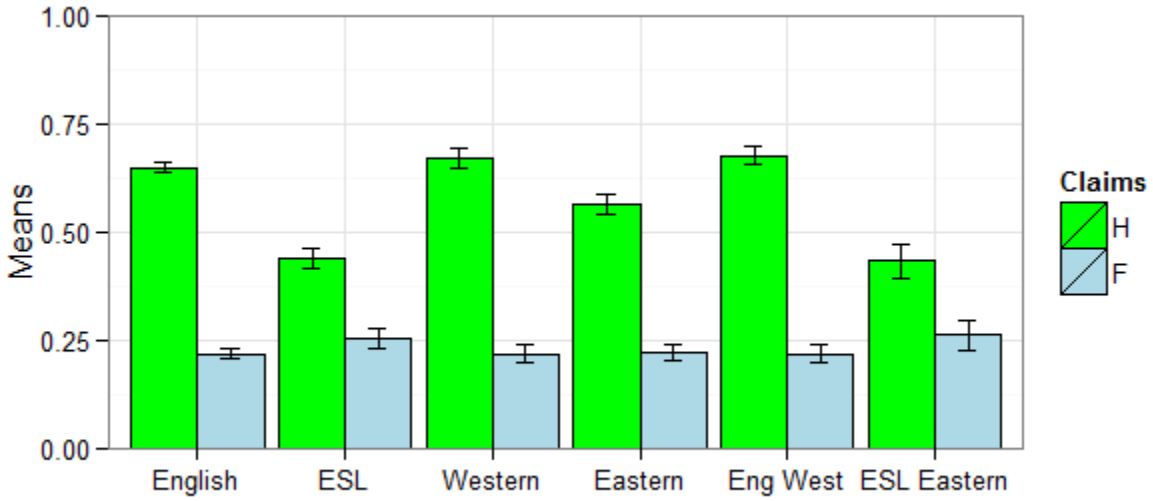


Figure 19. Study 3 Cultural invariance of Foil claiming.

A comparison of H (Reals claiming) and F (Foils claiming) for groupings based on language ability (English / ESL), culture (Western / Eastern) and both (Eng West / ESL Eastern). Error bars show standard error.

group certainly does not claim fewer Foils (their difference is .04 [-.04, .12], $t(72.64) = 1.11$; $d = 0.23$) contrary to assumptions about overclaiming capturing individualist-style self-enhancement.

4.4.3.5 Experience and Confidence with English (Self-Report)

Before any ability or personality measures, respondents were asked their experience and confidence reading and writing English (on a scale of 0 to 100). Unsurprisingly, the two measures correlated strongly, $r(336) = .88$*** [.85, .90], and both followed age of learning English: $r(336) = -.70$*** [-.75, -.64] and $r(336) = -.60$*** [-.67, -.53], respectively. Likewise for years in Western countries: $r(333) = .62$*** [.54, .68] and $r(333) = .52$*** [.44, .60], respectively.

It is worth noting that 61% chose the maximum 100 for experience, and 45% chose 100 for confidence. Excluding those extremes yielded *M* (*SD*) of 78.74 (16.85) and 78.55 (17.24) respectively.

50% of respondents rated their experience and confidence exactly the same, and 14% rated their confidence higher than their experience. The difference score (a kind of over-confidence) correlates with confidence ($r(336) = .46$*** [.37, .54]) but not experience or any other demographic variable (although marginally with Western ethnicity), and is apparently warranted, since it predicts Accuracy and Word score ($r(336) = .14$* [.03, .24]). It also, interestingly, predicts self-deceptive enhancement ($r(336) = .13$* [.02, .23]).

Overall, the parallel of self-rated expertise with both A and B supports independent research (Atir, Rosenzweig, & Dunning, 2015), showing that expertise, actual or believed, increases both discriminant claiming (Accuracy) and indiscriminant claiming (Bias).

4.4.3.6 Accountability

Comparing Study 2 (prescreen) claiming with Study 3 claiming of the same items we found enough of an increase in Foil claiming to also increase Bias and marginally lower Accuracy (see values in Test-Retest table, above). For all measures, the difference was greater for the lab condition than the online condition, indicating that claiming (particularly for Foils) increased with accountability. This is contrary to previous research (Paulhus et al., 2003), in which increased accountability decreased Bias. It may be that, since VOCE seems more like an ability test than a common knowledge survey, increased effort from students induces more overclaiming.

4.4.3.7 UBC Word Test and Personality Measures

These measures were all normally distributed. The correlation matrix in Table 6 shows their inter-relationships.

	Word	**HH**	**SDE**	**IM**	**NPI**
UBC Word Test	---	.20**	-.06	-.02	-.19***
Honesty-Humility	.20**	---	.11	.51***	-.34***
Self-Deceptive Enhancement	-.06	.11	---	.28***	.31***
Impression Management	-.02	.51***	.28***	---	-.22***
Narcissism	-.19***	-.34***	.31***	-.22***	---

Table 6. Study 3 correlations between Word Test and Personality Measures.

***Note.* *$p < .05$, **$p < .01$, ***$p < .001$.**

The strongest correlation shown, between HH and IM, suggests that the Honesty-Humility measure is susceptible to faking (or that this measure of Impression Management depends on

one's honesty). Interestingly, HH also follows Word score (and VOCE Accuracy) suggesting a verbal ability correlate, while Narcissism shows the opposite pattern.

4.4.3.7.1 Structural Validity

Loevinger, (1957) argued that scales should have good structural validity, i.e. that the scale's internal structure parallels the external structure of a target trait. In this situation, it makes sense to compare how VOCE item performance (item-total correlation with overall VOCE Accuracy, IT.A) compares with the same subject's score on an independent vocabulary test, in this case, the UBC Word test. For the Study 3 item set, correlating IT.A with their IT.W (how well the item's claim rate follows UBC Word score), we find for Reals, $r(44) = .90$*** [.82, .94], and for Foils, $r(48) = .89$*** [.82, .94], suggesting that ability within VOCE strongly parallels ability as measured by the UBC Word test, despite them testing different words in different ways.

4.4.3.8 Follow-up Items

At the end of the survey, there were two follow-up items, asking subjects what percentage of VOCE items they thought were fake, and what percentage they'd seen before in any other study.

On average, subjects reported they thought almost half the items weren't genuine (43.01 (21.65)), which is close to reality. Negative correlations with Foil claiming and overall Bias suggest this perception suppressed overclaiming yet did not significantly affect Accuracy, which is consistent with previous research indicating that warning subjects about the existence of Foils lowers claiming but retains the difference (Paulhus et al., 2003).

It is interesting to note, however, that this perception of fakes also correlated negatively with Word score ($r(110) = -.20$* [-.37, -.02]) and Impression Management ($r(110) = -.19$* [-.36, -.00]) hinting at an ability or personality bias in suspicion of fakes.

On average, subjects estimated they had seen about a third of the items before (32.41 (28.94)), yet this self-report item showed no correlation with actual exposure (i.e. having matching prescreen HSP IDs), $r(110) = -.01$ [-.19, .18]. Curiously, this item marginally followed Foil claiming in general ($r(110) = .18$ [-.01, .35]), suggesting that people who thought they'd seen the test before also thought items were things they were familiar with, regardless of actual exposure. Respondents who actually had seen items before showed reduced Foil claiming (similarly on Repeat items, $r(336) = -.15^{**}$ [-.26, -.05], and New items, $r(336) = -.14^{**}$ [-.24, -.03]), but this may be more about the kind of person who accurately recalls their HSP ID (see results below on HSP ID matching). Altogether this suggests an important part of Foil claiming is memory bias, as seen in previous research (Paulhus et al., 2003).

Fortunately, these two follow-up items did not correlate ($r(110) = -.07$ [-.26, .11]), so it is unlikely that their presence on the same page of the survey influenced each other, or that they were capturing some kind of skepticism response bias.

4.4.3.9 Confounds Introduced by HSP ID Matching

For linking these results with Study 2 results, we relied on an HSP ID code that respondents needed to remember how to construct (from family birthdates) from when they did the prescreen (from days to weeks previously). Those who produced correctly matching IDs were not necessarily a random sample, however, which may affect results. Students who entered IDs that matched with prescreen results had significant correlations with VOCE Accuracy ($r(336) = .15^{**}$ [.04, .25]), UBC Word test score ($r(336) = .16^{**}$ [.06, .27]), experience with English ($r(336) = .13^{*}$ [.02, .23]), Foil claiming ($r(336) = -.15^{**}$ [-.25, -.05]), and Narcissism ($r(336) = -.12^{*}$ [-.22, -.01]), hinting at a competence confound.

4.4.4 Predictors of Item Behavior

4.4.4.1 Reals

Unlike Study 2, Reals claim rates did significantly correlate with the Difficulty of the word, $r(44) = -.40^{**}$ [-.62, -.12]. The change in significance is likely due to their being double the number of items in this set; the relationship is not significant with either the Repeat or New set alone.

4.4.4.2 Foils

For strictly lexical measures, no significant predictors were found for Foil claiming or performance.

In analyzing each study, predictions from ELP measures were inconclusive or inconsistent. However, this may be due the limited power of testing only the small set of items from a single study. The next chapter considers the broader picture over all our VOCE work and finds more convincing patterns.

4.5 Discussion

The purpose of this study was to confirm and validate VOCE as an overclaiming instrument. To that end we had several hypotheses.

Test-retest reliability wasn't as high as one might hope, although surprisingly, concurrent reliability was higher. The test-retest difference may be explained by the change in context, so a more controlled study is called for. Concurrent reliability was remarkable high, but that was after four problematic items were removed, which indicates the importance (and potential) of having good models for selecting items.

4.5.1 H1: Better distributions H and F

Tweaking the difficulty of the New set worked exactly as intended, producing a well-centered distribution of Reals claiming with no ceiling effects. The few negatively performing Reals suggest that Reals claiming must be carefully tailored to the sample: too hard Reals act like Foils, and too easy Reals create ceiling effect problems noted in Study 1.

Foils claiming was also improved, which may be due to skill or luck in using WordGen, and possibly also due to increased accountability. The Foils claiming distribution achieved in Study 3 may be about as even as one can reasonably get.

4.5.2 H2: Convergent Validity - Capturing Language Ability

The UBC Word test measured vocabulary ability in a different way with different words, yet those scores correlated highly with VOCE Accuracy. Using both VOCE measures (either A & B or H & F) to predict Word score produces an R^2 of .46, $F(2,335) = 144.78$, $p < .001$. It worth noting that administering VOCE also takes considerably less time and yields more information.

4.5.3 H3: Cultural Invariance - Replication of Study 1 findings

While this was an unexpected finding in this course of research, it is notable that this pattern was reproducible. Since this questions some assumptions about overclaiming, self-enhancement, and possibly understanding of cultural differences, further research is warranted.

4.5.4 H4: Personality Correlates

While VOCE did correlate in expected ways with Narcissism, it did not do so strongly, and it failed to adequately capture SDE or IM. It may be that vocabulary overclaiming does not trigger the same kind of ego identification that some other domains do. The use of vocabulary may make it a more efficient ability measure, but also a less efficient personality measure.

Interesting, though, was finding that HH followed Accuracy. Perhaps VOCE may capture personality differently than other overclaiming scales.

4.5.5 H5: Accountability

While VOCE did not react to contextual changes in accountability as expected, it still did so in an understandable way. This makes it more suitable for higher-stakes testing, where pressure to perform will induce more overclaiming. VOCE may prove useful in crucial ability tests, particularly to identify risky tendencies to claim unwarranted ability.

4.5.6 H6: Previous Exposure

Finding that students cannot accurately recall whether they've seen the test before is perhaps not surprising, but it is telling that this false-recall tendency generalized across items. This finding also suggests VOCE may be useful in predicting risky tendencies in performance estimates. The confounds introduced by matching via HSP IDs needs to be addressed, however, and further research would be needed to clarify.

4.6 Conclusions

Study 3 provided broader support for the utility of the VOCE. It can work well as a vocabulary measure and our techniques for item selection are reproducible. In the next section, we discuss improvements in item selection gleaned from all the studies combined.

Chapter 5: Psycholinguistic Predictors of Item Behavior

Over the course of the three studies, we were interested in discovering ways of predicting item performance *a priori* using psycholinguistic measures. Each study's results gave us a glimpse, but an incomplete and sometimes contradictory image. Hence, did not report those analyses in each study. Like the parable of blind men encountering an elephant at different points (trunk, ears, legs), small samples may lead to misleading and inconsistent conclusions. After three studies, experimenting with different item selection techniques (largely based on incomplete versions of the models described below), a more consistent and encompassing picture emerges.

From these studies (and other pilot studies), we gathered information on 171 VOCE items. Because we tried different approaches, we generated enough variance to find some patterns. The following tables show correlations between psycholinguistic measures (from the English Lexicon Project database) and VOCE item behavior.

We have considered basically two kinds of psycholinguistic measures: *lexical measures* which are based on the word as a letter sequence (in reference to the corpus (body) of letter sequences that constitute English vocabulary), and *behavioral measures*. Semantic measures were not considered here due to complexity, and also because they would be useless for nonwords.

The behavioral measures come from two common kinds of psycholinguistic experimental techniques: the lexical decision task (LDT) and the naming task. The LDT is a reaction time (RT) test to distinguish genuine words from nonwords, and it produces both RT latency and accuracy scores. The naming task requires participants to correctly name a given word as quickly and accurately as possible. Measures from both these tests are included in the ELP database we

referenced and are considered “the gold standard in developing computational models of lexical processing” (Balota et al., 2007).

Although several measures were considered, only the ones that proved useful are reported here. To assist the reader, I first present a handy glossary of variables used from the ELP database.

5.1 Lexical Measures

Length: The number of letters in the word.

Ortho_N: An orthographic neighbor of a word is one that is identical except for one letter change, for example, “dot” and “cog” are orthographic neighbors of “dog”. Ortho_N is the number of neighbors a word has. For nonwords, this is one measure of how English-like it looks.

OLD20: Orthographic Levenshtein Distance (OLD) is the number of letter changes required to turn one word into another, and thus measures the “distance” between any two letter sequences, whether words or nonwords. OLD20 is the average OLD to the 20 closest genuine words; lower numbers indicate a denser “neighborhood”. This measure is empirically validated (Yarkoni, Balota, & Yap, 2008) and popular in linguistics research as a measure of orthographic similarity, or basically how alike words are in terms of their letter sequence. This is another way to assess how close a nonword is to genuine English words.

Log_Freq_HAL: Frequency norms from the Hyperspace Analogue to Language (HAL) corpus are based on a collection of roughly 131 million words gathered from 3,000 Usenet groups during February of 1995 (Lund & Burgess, 1996). These occurrence counts were log-transformed to act as a linear index of how frequently a word is used in typical English, and so only applies to genuine words.

5.2 Behavioral Measures

I_Mean_Accuracy: This is the proportion of accurate responses (in a lexical decision task) for a particular genuine word. It represents how easily the word is recognized as genuine.

NWI_Mean_Accuracy: This is the corresponding behavioral measure for identifying nonwords in a lexical decision task.

I_NMG_Mean_Accuracy: This is the proportion of accurate responses (in a naming task) for a particular genuine word.

NWI_Zscore: This is the standardized RT latency for nonwords in a lexical decision task.

5.3 Predicting Item Behavior

The two criteria of item behavior being modeled here are 1) item claim rate, and 2) item performance (item-total correlation with Accuracy). These were averaged over whatever surveys the items appeared in (including some smaller pilot studies not mentioned in this paper). Correlations were averaged via Fisher *z* transformations.

None of the predictors mentioned here appeared to have non-linear relationships with either criterion. Q-Q plots of the regression models typically showed moderate tails, but there were no influential outliers with Cook's distance greater than 0.5. Surprisingly, no interactions were found (that weren't explained by influential outliers). Models were constructed using a backwards deletion approach, starting with significant correlations. Considering that there are no existing attempts to predict overclaiming item behavior, all the models here are exploratory.

5.3.1 Predictors for Reals

In Study 1, Reals were too easy: A ceiling effect distorted results so much that better predictions were found in the Good subset of less than half the items. Study 2 Reals (more difficult) performed fairly well, and in the Study 3 New set (chosen to be even more difficult), a

few were so unusual as to elicit negative item-total correlations. Obviously, then, the sweet spot is in between, and having gathered data at the extremes allowed us to construct a stronger model.

Predicting Claiming of Reals

We found three ELP measures that strongly predict claim rates for Reals (Table 7), with an R^2 of .90, F(3,61) = 179.15, $p < .001$. No interactions were found among these predictors.

ELP Measure	***β***
I_Mean_Accuracy	.49, t(61) = 10.10, $p < .001$
Log_Freq_HAL	.30, t(61) = 6.47, $p < .001$
I_NMG_Mean_Accuracy	.19, t(61) = 3.99, $p < .001$

Table 7. ELP variables predicting Real claiming.

This model is fairly straightforward. The tendency to claim knowledge of a word is predicted by how accurately it is recognized as a real word, how frequently it appears in print, and how readily it is spoken. The obviousness of these predictors explains the large amount of variance explained. It is worth noting that the Difficulty ranking from the Top 1000 list contributes nothing to this model, so future research will ignore that.

5.3.1.1 Predicting Performance of Reals

Predicting Reals performance was somewhat more challenging, but the model shown in Table 8 produced an R^2 of .36, F(3,61) = 11.25, $p < .001$. No interactions were found among these predictors.

Here we see two lexical predictors and one behavioral predictor that contribute to a word's discriminability in a vocabulary test. The lexical predictors are telling: OLD20 indexes orthographic distance from other words – a measure of how lexically unconventional it appears. For example, "quagmire" and "pellucid" have greater OLD20 scores than "attitude" or

"contract" do. But OLD20 also increases with Length, so together they capture words that are difficult but not just because they're long. Apparently it's the short, weird words that make the difference in a vocabulary test.

ELP Measure	***β***
Length	-.64, t(61) = -4.95, $p < .001$
OLD20	.40, t(61) = 3.07, $p < .01$
I_NMG_Mean_Accuracy	.35, t(61) = 3.87, $p < .001$

Table 8. ELP variables predicting performance of Reals.

These models suggest that optimal Reals for VOCE could be first selected to have claim rates appropriate to the sample, to avoid ceiling or floor effects. Within that, Reals can be selected to provide maximal performance for efficient discrimination of ability.

5.3.2 Predictors for Foils

An understanding of Foil behavior proved to be more elusive. Without any clear signals from any one study, we tried three different techniques for generating Foils: using Wuggy and WordGen software, and by borrowing from the ELP database. In every case, however, we had to guess at what might yield better foils. It turns out the variability resulting from guessing was fortuitous, because it created enough variance to show us the error of our ways.

In Table 9, consider the Ortho_N and OLD20 lexical measures. They are similar in that they are based on number of letter changes between fake and genuine words, but Ortho_N captures proximity and OLD20 measures distance, so they are inversely related ($r(83) = -.82$*** [-.88, -.74]).

Predictor	Overall Foil Claiming	Overall Foil Performance
Ortho_N	.30**	-.32**
OLD20	-.32**	.17
NWI_Zscore	-.71***	-.33*
NWI_Mean_Accuracy	-.72***	-.42**

Table 9. Correlations of select ELP measures with Foil claiming and performance.

Note that they predict both claiming and performance, but in opposite directions! Whatever made Foils more likely to be claimed also reduced performance. Previous attempts at choosing optimal Foils (typically by selecting for higher Ortho_N or lower OLD20) was confounded by this paradox. This revelation (only available after analyzing all three studies together) is a game-changer: When selecting Foils from ELP based on higher Ortho_N (because that did predict claiming), there too many common suffixes. Allowing for Ortho_N of zero (which many of our best performing Foils had) opens up thousands of possible nonwords to consider from ELP. This, in turn, allows us to use their behavioral measures (the NWI variables in the table): They are much stronger predictors of claiming and performance, and thankfully, in the same direction!

5.3.2.1 Predicting Foil Claiming

With the ELP behavioral data, we were able to predict claim rates for Foils, with an impressive R^2 of .60, $F(2,35) = 26.35$, $p < .001$ (Table 10). No interactions were found among these predictors.

This success should not be surprising. After all, the first measure (NWI_Zscore) is of relative reaction time and the second is of accuracy, both for the lexical decision task of identifying fake words. Only these behavioral measures retain significance in a regression model,

which means our best bet for predicting Foil claiming is to stick to nonwords from the ELP database.

ELP Measure	***β***
NWI_Zscore	-.29, t(35) = -2.60, $p < .05$
NWI_Mean_Accuracy	-.32, t(35) = -2.86, $p < .01$

Table 10. ELP variables predicting Foil claiming.

5.3.2.2 Predicting Foil Performance

Performance for Foils could also be reasonably modeled (Table 11), with an R^2 of .37, F(3,34) = 6.75, $p < .01$. No interactions were found among these predictors. A closer look indicates that the model is curious for a number of reasons. First, both lexical predictors (Ortho_N and OLD20) contribute significantly, even though they correlate highly. This differentiation may derive from the fact that Ortho_N is a fairly crude measure, being largely single-digit positive integers. There is much variability of (continuous) OLD20 within a single value of Ortho_N. Second, the lexical predictors are negatively related but both contribute in the same direction! This may be a byproduct of naive Foil selection using three different techniques. Selecting for one may have restricted the other. Future research can be more aware of their contrary influence on claiming and performance. Finally, the behavioral measure is the weakest predictor! This pattern validates a hunch of mine that drove this line of research, that something

ELP Measure	***β***
Ortho_N	-.62, t(34) = -2.43, $p < .05$
OLD20	-1.54, t(34) = -3.28, $p < .01$
NWI_Mean_Accuracy	-.58, t(34) = -3.49, $p < .01$

Table 11. ELP variables predicting performance of Foils.

about the construction of Foils could make them more or less useful in overclaiming. Of course, the effect was opposite to my prediction, but such is research.

5.4 Discussion

With no previous research into modeling overclaiming item behavior, the mere fact that substantial amounts of variance can be explained at all, for both Reals and Foils, both claiming and performance, is encouraging. Future research will no doubt refine these models, but this exploratory research begins carving a new path where there was none before.

An interesting pattern shown in the models is that claiming is largely predicted by behavioral measures whereas performance has important lexical predictors. The ELP behavioral measures, since they are based on a lexical decision task, have obvious face validity: How well people do distinguishing genuine words from nonwords should naturally contribute to how likely they are to claim familiarity. The fascinating discovery is that lexical properties, particularly OLD20, can be strong contributors to how well both Reals and Foils perform.

Chapter 6: Conclusions

Together, the studies presented here permit several overall conclusions. One is that the new VOCE instrument is a useful measure of vocabulary. As such, it taps a less volatile knowledge base than previous overclaiming measures. Broader conclusions included key recommendations about overclaiming in general and item selection in particular.

6.1 Selection of Reals does Matter

Study 1 showed how Reals that are too easily claimed can alter behavior of the instrument and change the relationship between Accuracy and Bias (and possibly anything else they might correlate with). With mostly easy Reals, respondents' hit rates were near maximum so variance came more from Foil claiming. This, in turn, skewed Accuracy, narrowed Bias, reversed their relationship, and dampened their predictions of other measures. The so-called 'Good' item subset was more useful than the entire set, despite being less than half the size. The value of adjusting Reals' difficulty was confirmed in Study 2. I am not aware of any overclaiming research that identifies this potential hazard. Clearly, ceiling effects in Reals need to be avoided in developing overclaiming measures.

Study 3 showed that Reals can also cause problems if too difficult. If no one in the sample genuinely recognizes a Real, it acts like a Foil and contaminates Accuracy scores. The Difficulty ranking from the Vocabulary.com Top 1000 list (supposedly developed using Item Response Theory over a large sample) was only a crude starting point, and not precise enough to adequately match the ability found in our sample (i.e. no significant correlation between that Difficulty ranking and claim rates for the Repeat or New set). Fortunately, the model built from ELP data was much more precise.

Our regression model for Reals performance shows that simply matching difficulty to ability is not enough. Knowing that short, weird words perform better is crucial in building an efficient instrument.

While the overclaiming technique can give useful results in even simple ad hoc designs, it is important to realize that for Accuracy to be a meaningful ability measure, selecting Reals should be taken as seriously as selecting items on any ability test. Indeed, the value of an overclaiming Accuracy measure depends critically on the ability of Real items to discriminate ability.

6.2 Foils Are Complicated

The nature of Foil claiming appears to be very different than that of Reals. For one thing, Foil claiming is not a normally-distributed phenomenon. It appears to behave more like count data and have something like a Poisson distribution. The latter, by definition, have means equal to their variances. Without going into detail, the means of the Foil distributions have tended to correlate highly with their variances (not the case for Reals.)

The tradition of choosing Foils intuitively may be hazardous. Only after gathering several studies worth of data did it become clear that predictors for Foil claiming and Foil performance sometimes pointed in opposite directions. Perhaps that was part of the reason that the paired Foils and Reals in Study 1 behaved in opposition. Although Real performance is important, some degree of Foil claiming is also necessary: Otherwise, Accuracy scores are merely a self-report of claiming, which is of limited value (but see Ackerman & Ellingsen, 2014). Any study using overclaiming should report the proportion of subjects that claimed no Foils, a relevant issue in interpreting results.

6.3 Cultural Invariance

An unexpected finding was that Foil claiming seemed unaffected by ethnicity. This is at odds with the conception of overclaiming as egocentric self-enhancement, since that should vary between the individualist and collectivist cultural samples we had. The finding appears robust, given that it replicated with a different sample and with different items. Clearly more research would help elucidate this anomaly.

6.4 VOCE May Not Tap Traditional Overclaiming Bias

Another surprise was that VOCE didn't capture Self-Deceptive Enhancement or Impression Management. Perhaps this is in line with it not capturing cultural differences. Because overclaiming is known to depend on the respondent's valuing or identification with the knowledge domain, perhaps vocabulary doesn't engage ego mechanisms to the same extent that knowing book titles, science terms, political events or musical artist's names does. Vocabulary has many advantages as a general-purpose knowledge domain, but its neutrality may have the downside of being less engaging.

Nonetheless, OCT Bias was associated with narcissism and Honesty-Humility, so it retains value as a dual ability/personality measure. This trait-like property has the positive potential of being immune to faking good or bad, although more research is needed to confirm this. It also raises the question of what other non-cognitive individual differences might be predicted by VOCE Bias. The VOCE version may be less about self-enhancement and more about difficulties in negotiating the boundaries between knowledge and ignorance, perhaps a kind of meta-cognitive skill. Isolating and measuring this trait may be useful in predicting a variety of behaviors, predictions impossible with self-reports. Our recent research (Paulhus & Dubois,

2014), for example, indicated that high Bias tendencies reflect poor ability calibration, which undermines academic performance (Paulhus & Dubois, 2014).

6.5 Future Directions

Even before fully identifying the psycholinguistic prediction models in Chapter 4, our item selection techniques had evolved enough to produce a solidly-performing VOCE instrument. Of course, further research is necessary to thoroughly test these models on diverse samples, but our results provide a firm foundation for future custom generation of novel item sets tailored to specific populations.

Most important to future research on the VOCE will be the application of Item Response Theory (IRT). That technique permits detailed analysis of item performance across the entire latent factor (or factors) underlying vocabulary ability and overclaiming. With separate models for both claiming and performance, a computer adaptive testing application could, conceivably, dynamically select items first to optimize claiming (to avoid ceiling and floor effects for the sample being evaluated), and then winnow items to optimize performance. In this fashion, IRT should facilitate the building of a large item pool useful for linking results across samples with differential talent, thus facilitating tailored testing techniques.

This work may also contribute to, and be informed by, ongoing discoveries in psycholinguistics. For example, word length was found to have a U-shaped pattern of influence on lexical decision tasks, (with 5-8 letters being the bottom of the U), which may help refine our models (New, Ferrand, Pallier, & Brysbaert, 2006). ELP researchers have recently identified individual differences in visual word recognition (Yap, Balota, Sibley, & Ratcliff, 2012), which our work could help elaborate.

As robust as the overclaiming technique has been, the failure to tease apart Real and Foil performance has undermined exploration of the nature of overclaiming itself. With more powerful tools developed using the kind of research described here, perhaps we can unlock some mysteries of the "unknown knowns".

References

Ackerman, P. L. (2000). Domain-Specific Knowledge as the "Dark Matter" of Adult Intelligence Gf/Gc, Personality and Interest Correlates. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *55*(2), P69–P84. http://doi.org/10.1093/geronb/55.2.P69

Amati, F., Oh, H., Kwan, V. S. Y., Jordan, K., & Keenan, J. P. (2010). Overclaiming and the medial prefrontal cortex: A transcranial magnetic stimulation study. *Cognitive Neuroscience*, *1*(4), 268–276. http://doi.org/10.1080/17588928.2010.493971

Anderson, C. D., Warner, J. L., & Spencer, C. C. (1984). Inflation bias in self-assessment examinations: Implications for valid employee selection. *Journal of Applied Psychology*, *69*(4), 574–580. http://doi.org/10.1037/0021-9010.69.4.574

Ashton, M. C., & Lee, K. (2009). The HEXACO–60: A Short Measure of the Major Dimensions of Personality. *Journal of Personality Assessment*, *91*(4), 340–345. http://doi.org/10.1080/00223890902935878

Atir, S., Rosenzweig, E., & Dunning, D. (2015). When Knowledge Knows No Bounds Self-Perceived Expertise Predicts Claims of Impossible Knowledge. *Psychological Science*, *26*(8), 1295–1303. http://doi.org/10.1177/0956797615588195

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459.

Beer, J., & Hughes, B. (2012). Medial orbitofrontal cortex is associated with shifting decision thresholds. *Neuroimage*, 889-898.

Block, J. (1965). *The challenge of response sets: Unconfounding meaning, acquiescence, and social desirability in the MMPI*. East Norwalk, CT, US: Appleton-Century-Crofts.

Borgatta, E. F., & Corsini, R. J. (1960). The Quick Word Test. *The Journal of Educational Research*, *54*(1), 15–19. http://doi.org/10.1080/00220671.1960.10882672

de Vries, R. (in press). Overclaiming behavior: A matter of openness to experience, not Honesty-Humility. *Journal of Personality and Social Psychology*.

Duyck, W., Desmet, T., Verbeke, L. P. C., & Brysbaert, M. (2004). WordGen: A tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 488–499. http://doi.org/10.3758/BF03195595

Heine, S. J., & Lehman, D. R. (1997). The cultural construction of self-enhancement: An examination of group-serving biases. *Journal of Personality and Social Psychology*, *72*(6), 1268–1283. http://doi.org/10.1037/0022-3514.72.6.1268

Hirsch Jr., E. D., Kett, J. F., & Trefil, J. S. (1988). *Cultural Literacy: What Every American Needs to Know*. Vintage Books.

John, O. P., & Robins, R. W. (1994). Accuracy and bias in self-perception: Individual differences in self-enhancement and the role of narcissism. *Journal of Personality and Social Psychology*, *66*(1), 206–219. http://doi.org/10.1037/0022-3514.66.1.206

Joseph, J., Berry, K., & Deshpande, S. P. (2008). Impact of Emotional Intelligence and Other Factors on Perception of Ethical Behavior of Peers. *Journal of Business Ethics*, *89*(4), 539–546. http://doi.org/10.1007/s10551-008-0015-7

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627–633. http://doi.org/10.3758/BRM.42.3.627

Loevinger, J. (1957). OBJECTIVE TESTS AS INSTRUMENTS OF PSYCHOLOGICAL THEORY: Monograph Supplement 9. *Psychological Reports*, *3*(3), 635–694. http://doi.org/10.2466/pr0.1957.3.3.635

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208. http://doi.org/10.3758/BF03204766

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide.* New York: Cambridge University Press.

Marchman, V. A., & Fernald, A. (2008). Speed of word recognition and vocabulary knowledge in infancy predict cognitive and language outcomes in later childhood. *Developmental Science*, *11*(3), F9–F16. http://doi.org/10.1111/j.1467-7687.2008.00671.x

McCrae, R. R., & Costa, P. T. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology*, *51*(6), 882–888. http://doi.org/10.1037/0022-006X.51.6.882

Morris, E. (2014). *The Unknown Known*. Documentary.

New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review*, *13*(1), 45–52. http://doi.org/10.3758/BF03193811

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*(3), 598–609. http://doi.org/10.1037/0022-3514.46.3.598

Paulhus, D. L. (2011). Overclaiming on personality questionnaires. In M. Ziegler, C. MacCann, & R. D. Roberts, *New perspectives on faking in personality assessment* (pp. 151-164). New York: Oxford University Press.

Paulhus, D. L., & Dubois, P. J. (2014). Application of the Overclaiming Technique to Scholastic Assessment. *Educational and Psychological Measurement*, 0013164414536184. http://doi.org/10.1177/0013164414536184

Paulhus, D. L., & Harms, P. D. (2004). Measuring cognitive ability with the overclaiming technique. *Intelligence*, *32*(3), 297–314. http://doi.org/10.1016/j.intell.2004.02.001

Paulhus, D. L., Harms, P. D., Bruce, M. N., & Lysy, D. C. (2003). The over-claiming technique: Measuring self-enhancement independent of ability. *Journal of Personality and Social Psychology*, *84*(4), 890–904. http://doi.org/10.1037/0022-3514.84.4.890

Phillips, D. L., & Clancy, K. J. (1972). Some Effects of "Social Desirability" in Survey Studies. *American Journal of Sociology*, *77*(5), 921–940.

Randall, D. M., & Fernandes, M. F. (1991). The social desirability response bias in ethics research. *Journal of Business Ethics*, *10*(11), 805–817.

Raskin, R. N., & Hall, C. S. (1979). A narcissistic personality inventory. *Psychological Reports*, *45*(2), 590–590. http://doi.org/10.2466/pr0.1979.45.2.590

Roeder, S., & Paulhus, D. L. (2010). Application of overclaiming technique to consumer research. *American Psychological Association.* San Diego.

Stanovich, K. E., & West, R. F. (1989). Exposure to Print and Orthographic Processing. *Reading Research Quarterly*, *24*(4), 402–433. http://doi.org/10.2307/747605

Swami, V., Papanicolaou, A., & Furnham, A. (2011). Examining mental health literacy and its correlates using the overclaiming technique. *British Journal of Psychology*, *102*(3), 662–675. http://doi.org/10.1111/j.2044-8295.2011.02036.x

Swets, J. A., Tanner Jr., W. P., & Birdsall, T. G. (1961). Decision Processes In Perception. *Psychological Review*, *68*(5), 301–340. http://doi.org/10.1037/h0040547

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual Differences in Visual Word Recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology. Human Perception and Performance*, *38*(1), 53–79. http://doi.org/10.1037/a0024177

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971–979. http://doi.org/10.3758/PBR.15.5.971

Zimmer, B. (n.d.). *SCIENCE OF LEARNING*. Vocabulary. com. Available online at www. vocabulary. com/educator-edition/Vocabulary. com. Retrieved from http://www.vocabulary.vocabulary.com/educator-edition/Vocabulary.com%20-%20Science%20of%20Learning.pdf

Appendices

Appendix A Demographic Variables for Studies 1 and 2

The culture categories in Studies 1 and 2 were limited by the format of HSP prescreen surveys. We had no input into how participant demographics were measured.

Participant ethnicity was reported by choosing from a list: African Canadian, Caucasian, Choose not to say, East Asian, First Nations, Hispanic, Middle Eastern, Other, Southeast Asian. From that list, we took Caucasian as proxy for Western cultural heritage, and combined East Asian and Southeast Asian under the label Asian heritage. South Asian was not an option, so such participants fell under 'non-Caucasian'. Japanese was also not an option, but that that category could be computed as those who reported being born in Japan and had Japanese (and not English) as first language. The category of native English speaker was computed as those who reported English as a first language or who were speaking English by age 4.

Gender was not restricted to only male or female. All items had a "Decline to Answer" option, which created missing data. The Age field allowed free text, so non-numeric entries had to be discarded. As a rule, we took the conservative approach of list-wise deletion.

Appendix B Slider Response Styles

One interesting finding emerged from the use in Study 3 of a linear 0-100 graphic input format. This format allowed respondents to use a mouse to click anywhere on a line rather than choose a specific option from a set, as in typical Likert scales. Some input was always required (they had to click somewhere) and a numeric value denoting their position appeared on the side of the screen. Thus this slider method is arguably at least as easy as selecting one of, say, 5 or 7 options.

Interestingly, a disproportionate number of respondents chose multiples of 10 (e.g. 50 or 60), despite this input requiring extra effort to choose such a round number. If, for example, their first click registered 61, such respondents adjusted it to 60. This suggests a curious individual difference in interacting with such input, that is, a need to lower the resolution of their response.

It may well be that using a graphic sliding-scale 0-100 random-access input (clicking or touching a point on a line) for an item could yield interesting extra information unavailable with conventional Likert scales. However, full analysis of the impact of this response style would require partialing out the item content, which is beyond the scope of this paper.

Nonetheless, to give a simple example, for one item (what percentage of items were thought to be fake), results seemed normally distributed and those choosing multiples of 10 (54%) were not significantly correlated with the item itself (r(110) = -.08 [-.26, .11]). This rounding response bias correlated with VOCE Accuracy (*r*(110) = .21* [.02, .38]) and Honesty-Humility (*r*(110) = .22* [.04, .39]) even though the item itself did not. Perhaps conscientiousness is involved. Further research on this intriguing phenomenon seems warranted.