Revisiting Itemmetrics: Do Psychologists Need to Watch their Language?

by

Alexis R. Georgeson

B.A., The University of North Carolina at Chapel Hill, 2012

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Arts

in the
Department of Psychology
Faculty of Arts and Sciences

SIMON FRASER UNIVERSITY
Fall 2016

Approval

Name:	**Alexis R. Georgeson**
Degree:	**Master of Arts (Psychology)**
Title:	***Revisiting Itemmetrics: Do Psychologists Need to Watch their Language?***
Examining Committee:	**Chair:** Dr. Thomas Spalek Professor

Dr. Rachel T. Fouladi
Senior Supervisor
Associate Professor _____

Dr. Allen Thornton
Supervisor
Professor _____

Dr. Deborah Bandalos
External Examiner
Professor
Dept. of Graduate Psychology
James Madison University _____

Date Defended: December 13, 2016

Abstract

Despite a long tradition of studying psychometric properties of self-report questionnaires in psychology, the literature identifying specific linguistic features of questionnaire items is sparse. Moreover, it is unclear whether linguistic features affect all individuals similarly, or interact with individual characteristics. The present study offers a novel methodological contribution whereby a variety of linguistic features, based on the domains typically studied by linguists (i.e., morphology, syntax, semantics), are proposed. To demonstrate how these itemmetrics can be used empirically, we analyzed data from the Center for Epidemiological Studies-Depression (CES-D) scale. Additionally, we probed interactions between sex and English fluency and each of the features to examine whether there were differential effects depending on the individual. Our results suggest that certain features may impact responding and interact with individual characteristics. We argue that our findings necessitate a stronger focus on this area of research.

Keywords: Item Response Modelling; Itemmetrics; Item Properties; Item Wording; Individual Differences

Acknowledgements

I would like to acknowledge my supervisor, Dr. Rachel T. Fouladi, for her mentorship, support, and dedication throughout this process. I would also like to thank the research assistants who put in their time and effort to help out in the lab. In addition, I would like to thank my mom and dad, my brother, for their constant support. I would also like to thank my friends, Aaron Garcia, Camille Weinsheimer, and Ric Hohn for their advice and for always having an open ear for me to work out aspects of this project. Finally, I thank my cat, Felix, for always keeping my lap warm while writing this.

Table of Contents

Chapter 1

Introduction

1.1 Test Theory

Self-report questionnaires allow for the measurement of a multitude of psychological constructs. Without self-report measures, understanding an individual's attitudes, preferences, feelings, and behaviors would be quite a challenge, if not impossible. Typically, a self-report measure will have a set of item stems and a response scale, which may or may not vary across the items. Item stems are primarily linguistic and can vary from a single adjective to multiple sentences. The response scale might be dichotomous (e.g., yes/no, true/false, agree/disagree) or it might have more options (e.g., 3-point "Agree"/"Neutral"/"Disagree", 5-, 7-, or 9-point Likert scales which ask the participant to indicate the extent to which they agree or disagree). Some instruments even have 100-point response scales (e.g., visual analog scale), allowing for more variability in responses. A respondent will provide responses to the items on a questionnaire, and then these responses are summed or composited according to a compositing rule to give a total score. Compositing rules are most often unit-weighted, where each item is given the same weight in the composite. For tests which have been properly validated, the total or composite score can be considered as a representation or estimate of the level of a particular trait that the individual has.

In classical test theory terms, an observed total score on a questionnaire is the summation of an individual's "true score" of the trait or attribute being measured, plus error, as demonstrated in the following equation from De Boeck and Wilson (2004):

$$Y_{pu} = \tau_p + \varepsilon_{pu} \tag{1.1}$$

In this equation, Y_{pu} is the test score for an individual p on occasion u, τ is the true score for individual p, and ε is the error term for person p at time u. To make an analogy, a researcher designing a test strives to create a sort of "net," which prioritizes the capture of the true amount of the trait of interest while minimizing interference (i.e., "error") from other variables. There are a number of sources of variability that can impact the observed

scores. Of primary interest is the variability across individuals of the "true scores" – the amount of the trait or construct. This is somewhat obvious – if the trait of interest did not vary, researchers would have little reason to study it. Next, there are sources of variability which may affect the level of the target trait at the time of measurement. These sources are numerous and, together, account for the "error" within a classical test theory framework. Theoretically, an informative test maximizes the measurement of the target trait and minimizes the contamination from other sources. Because there are so many sources of error in comparison to the single target trait, creating a high-quality test is often challenging. It is therefore important for researchers to continuously evaluate potential sources of contamination in order for psychological science to be meaningful.

As the field of psychology has aged, the process for creating and evaluating new measures has been discussed and systematized. For example, the American Psychological Association now recommends a particular approach when creating a new measure, which we will briefly discuss (Kingston, Scheuring, & Kramer, 2013; Wendler & Burrus, 2013). First, the researcher must conceptually describe what is to be measured, as well as the ways in which this trait is related to other traits (and the tools to measure those traits). Kingston et al. (2013, p. 167) recommends that the test maker also consider what the test will be used for (i.e., prediction or description), how many scores will come out of the test, how the scores will be reported and to whom, how the test will be administered, and the stakes of the test. Then, through consultation with experts, the test maker creates an initial item pool. Each item in this item pool must then be inspected to ensure that it is not "confusing, unnecessarily difficult, or tricky" and that it is as concise as possible (Wendler & Burrus, 2013, p. 286). The items are revised and removed until the test maker is satisfied with their test pool and it is then administered to a group, the psychometric properties such as reliability and concurrent and divergent validity are evaluated, and the test is either further revised, or considered to be a useful measure for the trait of interest.

Throughout this process, the test maker evaluates the items of the test and makes decisions to the best of their judgment about how to measure the trait while reducing error, such as confounds. However, criteria which are related to the language used in test items are not evaluated in any systematic way, and are rarely revisited after the psychometric properties of the test have been validated. Indeed, as Loevinger (1957) describes:

> So far as the constitution of the [item] pool is concerned, selection of a single item automatically excludes many others which differ slightly from the chosen one. Thus several items are neither randomly nor independently selected. The term "universe of items," moreover, obscures the fact that between the presumably unlimited number of items representing a given content and the finite pool of items actually studied there involves an **idiosyncratic, non-reproducible process**, the process by which the given investigator or group of investigators constructs or selects items to represent that content. Although

> this process, the constitution of the pool of items, is a universal step in test construction, it has not been adequately recognized by test theory. (Loevinger, 1957, p. 658, emphasis added)

According to Loevinger, the process of constructing the items is not prescriptive, and is quite arbitrary. Occasionally, the test makers may give participants an opportunity to provide specific feedback on the test in the form of a comment. Even more rarely, cognitive interviews might be conducted to understand how the individuals are responding to the items. The item pool may be adjusted if certain items are found to be problematic. For example, instruments developed within a certain era are sometimes adjusted because their stimuli are no longer relevant or even interpretable to the current generation. However, in the vast majority of circumstances, the individual test items are finalized in an admittedly careful, but highly subjective manner, *a priori* to administering the test. This tradition creates the impression that grammatical and linguistic aspects of items are trivial, yet, one would be hard-pressed to find cumulative evidence to support this impression. If we are to consider the possibility that there are components of test items that have measurable effects on responses, there would then be little reason to rest assured that a high-level revision of the language in test items is sufficient. Either way, this simply is not an option within the present tradition.

Consider a self-report questionnaire written in English. Undoubtedly, the questionnaire is intended for individuals who are fluent in English. Yet, the population of English speakers is not homogeneous. There seems to be the assumption that being fluent in English is sufficient for the questionnaire to be meaningful, or to be meaningful *in the same way* for all individuals who are fluent in English. If the researchers suspect that an item is being interpreted differently based on other characteristics within the broad group of English speakers, they may then conduct an analysis to explore this, and then change or remove the item. In this scenario, the researcher is primarily interested in identifying the function of the item, but they are rarely interested in why the item might be functioning differently. Perhaps a way in which we may anticipate this issue *a priori* is by formulating a more intricate understanding of what aspects of items may be leading to differential responding.

1.2 Depressive Symptomology

Self-report instruments are used within the field of clinical psychology to study, assess, diagnose, treat, and/or monitor symptoms within individuals (e.g., Carlson, 2013). Major depressive disorder is recognized in the DSM-5 as having affective, behavioral, and somatic symptoms and approximately 7.6 percent of the US population (22.8 million people aged 12 and older) experienced moderate or severe depressive symptoms between 2009-2012 (Pratt & Brody, 2014). The screening and diagnosis of depression in adults in clinical settings relies on communication with the individual, which may be in the form of a reciprocal

conversation, a semi-structured interview, a questionnaire completed by the individual, etc. While communication also occurs in the context of physical health diagnoses, there is often no other form of information-gathering within the context of psychological diagnoses. For example, a physiological diagnosis might be made based on vitals, blood panels, as well as consulting with the client with each type of information receiving a different weight. In contrast, biological and physiological measures are typically not as useful as the communication with the client when making a psychological diagnosis.

A clinical interview conducted by a trained psychologist is considered to be the most reliable method of making a differential diagnosis for depression (Sharp, Williams, Rhyner, & Ilardi, 2013). However, the first contact an individual makes with the healthcare system is likely to be with a primary care provider or a family practitioner, rather than a mental health specialist. A primary care provider must assess the overall health of their patient within a limited time frame in order to make referrals to appropriate care providers. Short, self-report screening tools for depression and other mental health concerns are invaluable in this setting because the patient is able to complete them while waiting to see the provider. Then the provider, or another staff member, can use the total score to determine the best course of action (i.e., referral to a different provider, further evaluation). The screening tool does not make any diagnoses, it simply indicates areas requiring further attention.

Screening tools for depressive symptomology are used in research settings as a way to determine which participants to include or exclude in a study. A researcher may also collect these measures in order to later statistically control for depressive symptomatology in their analyses. A screening tool for depressive symptomology is useful when it over-estimates individuals who may be experiencing depressive symptomology frequently or severely enough to qualify for a diagnosis. In other words, in order to capture pool of individuals who would meet criteria for major depressive disorder, an instrument must err liberally and capture false-positives. The purpose is to filter out individuals who have a small likelihood of being depressed and to avoid false-negatives (i.e., individuals who experience high levels of depressive symptomology or would receive a diagnosis of depression in a different context, but are not detected as such by the instrument)(Weissmann, Sholomskas, Pottenger, Prusoff, & Locke, 1977). The Center for Epidemiological Studies-Depression scale (CES-D) has been shown to perform well at detecting individuals who are depressed, though its ability to differentiate lower severity levels of depressive symptomology is limited (Radloff, 1991).

These screening tools inherently use written language. A challenge that comes with using language is that language, by its very essence, is inexact. Humans are able to express the same idea in a near infinite number of ways, yet, whether two utterances are ever really "equivalent" is one that philosophy has yet to settle. Moreover, a more psychological question might be whether two individuals interpret the same utterance in the exact same way. Language can be divided into various parts and described, and this is one pursuit undertaken by the field of linguistics. However, psychologists have yet to rule out the

possibility that variations in these linguistic categories which occur in self-report instruments may be influencing responding. In other words, the research methodology and knowledge that has been accumulated in linguistics has yet to fully make its way over into mainstream psychology. The current study is in part an attempt to use some of the methodology and theory developed within linguistics to analyze self-report questionnaires.

Previous research has found that lower levels of English fluency are related to higher levels of endorsement of negative mood and/or depressive symptomology. Whether this finding is related to linguistic features of the items interacting with the linguistic background of the individual, or if it is related to response styles within a particular linguistic background or cultural background remains to be seen. This finding serves as inspiration for the current study, but it is also inspired by the idea that linguistic features of psychological instruments are important to consider regardless of linguistic background of the individuals. If a certain type of feature lead to higher or lower endorsement of depressive symptomology, this would certainly be an interesting result to ponder upon. In the worst-case scenario, a feature which had large effects on reported depressive symptomology could lead to incorrect conclusions about the respondent.

In sum, self-report instruments continue to be an essential part of psychological science, in research as well as in practice. Self-report questionnaires may be the best tools we have, but there are inherent challenges to ensuring the validity of what is being measured, given that the questionnaires are linguistic in nature, and the intention is to measure behavior. In order to improve our extant instruments, develop better instruments moving forward, and clarify aspects of test theory, a more robust and collective knowledge of linguistic features and the way in which they may affect responses is required. This project will attempt to set up one approach to uncovering the effects of a variety of linguistic features ranging from semantic content to structural properties of the item stems. While the instrument investigated in the current paper purports to measure depressive symptomology, this project can be thought of as a template for evaluating linguistic features which may be applied to other areas of psychology, since self-report instruments are pervasive throughout the field.

1.3 Elucidation of Terminology and Theoretical Posture

In order to lay the groundwork for the current endeavor, some common ground must be established. To establish a common language for this discussion, we are transparent and explicit about the ways in which these terms are used. Additionally, in order to provide the necessary justification for this investigation, the theoretical underpinnings will be discussed.

1.3.1 Terminology

At best, it would be quite ironic to neglect to attend to the language and terminology used within a paper about the unique variability contributed by language. At worst, it would be

a careless logical error in the posturing of the argument. Thus, here we will make clear the intended senses of terms which occur frequently. This is done in order to reduce confusion and to allow the reader to have greater clarity as they consider this work. I have chosen to provide this as a separate, highly structured section as opposed to an in-line discussion in order for the reader to easily refer back if needed. Some of the descriptions are purely for functional purposes, while others are more theoretical and conceptual in nature.

Item An 'item' refers to a single question or a statement on a questionnaire which has a single response. The item consists of both an *item stem*, which is the stimulus (e.g., a question, a statement, an adjective) and *item response* (e.g., response scales, multiple choice). Unless explicitly stated, we will be discussing linguistic-based items which are read by participants, as opposed to pictorial or oral items. Instructions or other parts of the test are assumed to be separate from the item stem and item response. Primarily, the interest will be constrained to closed-ended items where the participant chooses one response option and any exceptions will be explicit. Furthermore, the use of the term "item" can be assumed to be referring to the item stem.

Questionnaire Generally, we will use the terms *test, questionnaire, assessment, measure, scale, instrument, inventory* interchangeably to refer to a set of items which measures some psychological trait or attribute. It will be made clear when we are discussing a specific *type* of questionnaire or test, as opposed to questionnaires in general. Unless specified, we are referring only to the types of tests that are linguistically based.

Property and Feature To date, there has not been a disambiguation in psychology between *feature* and *property* when referring to an item stem. When discussing the literature, we will attempt to remain consistent with the authors' use of these terms, which may sometimes substitute *characteristic* or *attribute*. For our purposes, we consider a *property* to be an attribute or trait of language and a *feature* to be a specific operationalization, or coding, of that property. For example, *nouns* could be considered to be a property, while *count of nouns* and *proportion of nouns* in a sentence are two different operationalizations of this property, and thus, features.

Content and Content Validity The discussion of the content of a test in the psychological literature is most often found in co-occurrence with content validity. The classical definition of content validity is the extent to which the test stimuli are representative of the "universe of tasks," which the test claims to measure (Cronbach & Meehl, 1955). Some have expanded this definition to specify that content validity is also concerned with the intended use of the test (e.g., Ebel, 1956; Sireci, 1998). Cronbach (1946, p. 476) makes an important point that "it is never possible wholly to separate the content of an item

from its form" and goes on to define the form as "the form of the statement, the choice of responses offered, and the directions, since all of these are part of the situation to which the person reacts". Similarly, Jackson and Messick (1958) discuss that "the complimentary constructs of content and style have special relevance to the questionnaire items, where the response-evoking properties of the particular item *form* may contribute markedly to response variance above and beyond the contribution of *content*". As a follow up to Jackson and Messick (1958), Bentler, Jackson, and Messick (1971) provide a closer inspection of different factors of content, ultimately concluding that content is elusive.

It then appears to be the case in psychology that the concept of "content" is not unambiguously defined. To date, the literature seems to refer to content as *what* the item is asking, but not *how* it is asking it. In an attempt to extract a meaning from some of the work that has been done, we have a simplistic definition and some questions that remain to be addressed. Our definition considers "content" to refer to the semantic, interpretive impression that is expected to be extracted from the item. Similarly to how syntax and semantics are difficult to separate within linguistics, it would seem that an open question is how to define content and where to place form and structure with respect to content. This question is highly relevant to the current investigation.

Construct The trait, attribute, or quality we want to measure, approximated through language, and hypothesized to "exist" (in no particular form, for our purposes) while never being directly observable (Cronbach & Meehl, 1955; G. T. Smith, 2005). We opt to use this basic definition rather than argue for a more nuanced definition.

Model A model is considered to be simplification of some aspect of reality, such as a process, or, in the case of statistical models, the relevant predictors that are associated with some outcome. A model is not considered to be a replica of reality, and thus, all models are "wrong," in this sense. However, it is the case that some models are superior to other models. Namely, those that are based on careful theoretical development and supported by empirical evidence are often the most useful models.

Individual Characteristics Broadly, individual characteristics are aspects of an individual that can describe their demographic make up, or some dimensions of their identity. Race, ethnicity, nationality, gender, language, and sex are some examples. For this study, the individual characteristics considered are sex and English fluency. Note that there are many additional characterisitics, such as personality, psychiatric makeup, or cognitive ability that could be explored, but were not chosen for the present study.

1.3.2 Theoretical Perspective on Responding to Questionnaires

In order to understand why features may be of importance to consider, we must first lay out the process of responding to a questionnaire. In other words, what are the cognitive processes that must occur in order for a respondent to choose a response? Tourangeau and Rasinski (1988) proposed a Four-Stage Model, which is the most frequently cited model to explain how the task of responding to a self-report questionnaire occurs (Jobe, 2010). The proposed model has four discrete steps: (1) Interpretation of the item, (2) Retrieval of information from autobiographical or long-term memory, (3) Generation or estimation of an answer using decision-making processes, and (4) Formulation of a response.

Other models that have been proposed elaborate upon Tourangeau's four-stage model. Esposito and Jobe (1991) developed the survey interaction model to account for some of the shortcomings that were in the four-stage model. The authors' main critique of the four-stage model is that it fails to take the broader context of the questionnaire situation into account. That is, other physiological or biological processes, the way questions are asked, the broader contextual factors, and person-specific factors. Their model can be thought of as an application of Bronfenbrenner's ecological models about embedded contexts to the survey-taking situation.

Esposito and Jobe (1991)'s framework has three components: the questionnaire context, the participants, and the survey interaction model. The survey context can include many variables, and these authors attempt to define the contextual variables which they argue could influence the questionnaire responding process. We have selected some of the contextual variables from their model to discuss with respect to the present study. One note is that this model was originally presented at a conference for the census bureau. While we see it as highly salient to self-report instruments in general, there are some aspects of it that apply more to census-type questions.

1. **Setting** The location of the respondent when they are completing the questionnaire. We might expect an individual to respond differently to a questionnaire when they are responding from within their own home vs. responding in a lab setting vs. responding in a medical setting. Characteristics of the location, such as the security and the presence of observers, would fall under this variable.

2. **Timing** The time at which the respondent completes the instrument with respect to other events. These events may be at the state or national level, at a community level, or at the personal level. For the CES-D, it might be important to consider when the data were collected with respect to time periods in the semester which are known to be stressful for students, such as mid terms and finals. Seasonality might also be facet of timing.

3. **Respondent characteristics** These might be demographic characteristics, socio-economic status characteristics, organismic characteristics like health, psychophysiology, and other salient characteristics like appearance, personality, or intelligence, experiential characteristics, such as the researcher or respondent's interest in the subject matter.

4. **Administration mode** The format used for the test, such as paper and pencil vs. computer test.

5. **Response security** The level of confidentiality and anonymity. We would argue that the actual level of confidentiality and anonymity may be less important than the *perceived* level of anonymity or confidentiality. Ethical standards require the individual to consent to how their data will be used and protected, but it is possible that their perception differs. For example, a questionnaire might be completely anonymized, but it is not impossible that the individual still feels as if they are being "exposed" to some extent when responding to a questionnaire.

6. **Attributes of the survey instrument** The authors of the original framework suggest that the questionnaire itself can be thought of as an aspect of the context, an then broken down further. For example, item characteristics such as the "content and salience of the question" or "the desirability and sensitivity of the question content," the response format, and the length or complexity are all important to consider. In the present study, we aim to elaborate upon our understanding of the questionnaire as a context.

7. **Incentives** Any effort made on behalf of the researcher which could increase the respondent's willingness to participate in the survey. For our purposes and those of most researchers, this variable is likely to be less of an important focus as part of the context. The reason is that incentives or payment are scrutinized by ethics committees. In some sense, the aim of ethics committees when considering payment is to minimize the influence of incentives as a contextual variable. Article 3.1 in the Tri-Council Policy Statement: Ethical Conduct of Research Involving Humans states "Because incentives are used to encourage participation in a research project, they are an important consideration in assessing voluntariness. Where incentives are offered to participants, they should not be so large or attractive as to encourage reckless disregard of risks...This policy neither recommends nor discourages the use of incentives. The onus is on the researcher to justify to the Research Ethics Board the use of a particular model and the level of incentives" (NSERC, 2014, p. 27). While the focus of the TCPS is on ethics rather than the meaningfulness of research findings, it would seem that the ethics board would still be concerned about a study offering a large monetary reward for participation, regardless of the level of risk involved. It

could be the case, however, that participants who feel that the research is particularly important because of their own priorities or experiences may be incentivized, in some sense, just by knowing about the purpose of the study. For example, an individual who had suffered from depression might be more willing to participate in a study which studied depression (and put more effort into the tasks involved) than they would be about a study which studied a topic they did not have such a personal connection toward.

The model of the survey interaction process as described by Esposito and Jobe (1991) then involves 7 phases. (1) The interviewer (or, in our case, the researcher), and respondent are orientated within the context of the study. (2) The interviewer asks questions. In the case of many questionnaires, the interviewer is not doing the asking, the participant is reading the question. (3) The respondent processes the question and chooses an answer. (4) The interviewer processes and records the response. Again, this step does not occur in every setting and certainly not when the questionnaire is being administered on a computer or pencil and paper. (5) The interviewer and respondent reorient themselves and continue to the next item, and steps 2-5 continue. (6) The interview is concluded. (7) The interviewer reviews and adjusts the protocol.

We recommend several adaptations to this model and propose the following model:

1. The respondent is orientated within the multi-layered context of the study. At this point, we argue that the most relevant contextual variables are the physical setting, the timing, and participant characteristics.

2. The participant reads the item and comprehends the item's meaning.

3. The participant processes the item and considers their possible responses. They then choose an answer.

4. They record their response.

5. The participant then moves onto the next item.

6. The participant reorients themselves to the context. Our understanding of this step is that as the participant progresses through the questionnaire, their impression of the context changes slightly. In other words, before they begin the questionnaire, they might only know that they will go through a questionnaire. Once they read the first question, they begin to have an idea of what the questionnaire is about. Perhaps the first item asks about how often they depressive symptomology. They will then incorporate this new information into their orientation. Each additional item and the way they respond allows for a potential readjustment and reinterpretation of the context. This is an important elaboration on Esposito's model as well as Tourangeau's.

1.3.3 Theoretical Perspective on Validity

Our conception of validity in the present endeavor is based largely on Messick (1989, 1995). In particular, validity is

> "an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores and other modes of assessment. Validity is not a property of the test or assessment as such, but rather of the meaning of the test scores (Messick, 1995). These scores are a function not only of the items or stimulus conditions, but also of the persons responding as well as the context of the assessment." (p. 742)

Messick (1995) goes on to describe six aspects of construct validity. Of these, we consider the generalizability aspect of validity to be of particular relevance because of the focus on the generalizability of a score and/or interpretation across "population groups, settings, and tasks" (Cook & Campbell, 1979). In addition, content validity is pertinent (see section on content). The view that validation is an ongoing process is an important justification for our investigation. It is important to continuously evaluate the tests and measures which are used in psychology, especially those that are used frequently.

The current investigation is exploratory and, therefore, we will not attempt to argue the case that linguistic features of questionnaire items and their interactions with the linguistic background of the respondent is a threat to the *validity* of a score. However, it is reasonable to suggest the possibility that linguistic features (and how the participant interacts with these features) contributes construct-irrelevant variance as a justification for this study.

1.4 Broad Research Aims

The present investigation will focus on making four contributions to the literature. The first contribution is a review of the literature which examines item features, tracing the history of itemmetrics. Next, a collection of examples, rooted in linguistic theory, and descriptions of the various ways in which item stems can be categorized – in terms of item properties– will be provided. These properties can be represented, or coded, for analyses in various ways. The different ways of coding properties will be referred to as features. The item properties will be categorized under larger domains, again, based on linguistic theory. This categorization can theoretically be applied to any type of instrument, though here they will be applied to a depressive symptomology instrument.

The next aims potentially make both methodological and substantive contributions through analysis of empirical data. Using item level analysis, we will examine the effects of certain features on levels of reported depressive sympotomology to determine whether

certain features result in higher or lower endorsements. The last contribution is an exploration of these features and their interactions with individual characteristics to determine if their functioning is consistent regardless of linguistic background or sex of the participant. In examining the effect of item features on depressive symptomology, we will also attempt to single out the optimal coding scheme for a particular property. This will also provide an examination of the interrelations among features, and properties.

Chapter 2

Literature Review

2.1 Introduction and Criteria

In order to place the present investigation in relation to other knowledge in psychology, examining the literature in some detail is necessary. The literature exploring linguistic features within self-report questionnaire items is sparse. Therefore, an effort is made here to provide a thorough and detailed literature review which articulates the historical developments related to the topic at hand. The most relevant literature will be discussed in detail, while literature that is peripherally related will be summarized more briefly.

The primary focus of the literature review is methodological as opposed to integrative. In other words, instead of condensing the findings across studies to draw broader conclusions about linguistic features, many of the studies included will be discussed individually to draw conclusions about the rationale behind exploring item features and the approaches used. We chose this approach because the analyses and findings from the studies on item features will not inform the current project, due to the variability in subject areas. For example, it is difficult to make an argument about how the findings from several studies using an adjective checklist with a dichotomous outcome variable may apply to a measure of depressive symptomology with four response options. The topic of linguistic features in questionnaire items is simply not saturated enough at this point to reveal any broad generalizations.

Thus, the historical progression, theoretical motivations, and methodological approaches are of importance to this study. The literature review will explore the initial motivations for studying item features, the types of features which have been of interest to researchers, the methodology used to study features, and the analytic approaches used.

Individual characteristics are another area of interest in the present investigation. The literature on differences in depressive symptomology across sex and linguistic background is much more dense in comparison to the literature on linguistic features. As such, the

literature review will simply summarize the general findings in these areas. The specific criteria that were used to guide the literature review were as follows:

1. Subject area and Methodology: The text must be ostensibly related to questionnaires, assessments, tests, or measurement via self-report in psychology.

2. Subject area: The text is taken from the literature on personality, mood and well-being, or education. Articles from industrial/organizational psychology or on specific clinical diagnoses (other than depression or mood disorders) were excluded.

3. Source: The text is either published in a journal or a book. Conference proceedings, presentations, or posters, dissertations, or unpublished work will largely be excluded unless the text was available online.

4. Approach: Empirical articles using quantitative methodology are prioritized since they relate to the present study. Texts with a theoretical focus will also be included as these are also relevant to the present investigation. Literature reviews and meta-analyses will be included on a case-by-case basis. Qualitative and other articles will be included if they can do any of the following: inform methodology for examining items, provide insight for interpreting items, provide insight about individual characteristics (namely language background), or provide theoretical motivations for studying item features.

5. Methodology: If the purpose of the text is empirical, the text must either conduct item level analyses or place emphasis on individual characteristics.

6. Texts on individual characteristics must be directly related to testing and assessment, depressive/negative symptomology, or linguistics.

2.2 Item Properties and Features

The extant research on item properties has largely fallen into two categories: (a) personality psychology and (b) educational testing. The research from personality psychology is most similar to the current study in purpose and execution. The education literature has provided some investigation into properties of test items, especially those that could be modified in order to manipulate difficulty. A significant contribution from the education literature is the concept and measurement of readability, which will therefore be discussed in some detail. Outside of personality and education research, but still within psychology, there has been some consideration of item properties, though this is not an area that has received much attention. The review will focus on research on self-report measures, especially those relating to health, mood, and well-being.

Each subsection will proceed by describing each area of the literature and discuss the rationale to studying item features, the approaches used, and relevant findings. We begin by discussing the motivations for exploring item features in personality and providing a historical overview. Again, methodology is of particular importance for the current study. Each section will proceed in an approximately chronological manner. Finally, the contribution and shortcomings of previous research will be discussed. When studies have also looked at individual characteristics interacting with item features, this will be discussed as well.

2.2.1 Personality: Initial Motivations

Researchers within personality psychology have had an interest in item content and item features on self-report measures for many years. Personality psychology has a long tradition of rigorous methodology. A thorough search of the literature reveals that, to date, personality psychology has provided the most focused study of item features within all the areas of psychology. A potential explanation for the attention given to item features is that personality theory from the beginning of the 20th century relied on certain assumptions about the behavior and measurement of traits (Allport & Odbert, 1936; Goldberg, 1963). Broadly, this section will discuss the motivations for studying item features. These motivations include researchers' attempts to understand and deter item instability, which resulted in targeted explorations of the feature of ambiguity. Looking at a slightly different angle, response sets or response styles were another concern in personality psychology which stimulated an interest in both understanding item properties as well as manipulating the items.

Item Stability

Early on in the history of self-report instruments in personality psychology, the description and measurement of item properties was a topic of interest. Researchers began to concentrate on item features in the personality literature in the 1930's as a potential route of improving the stability of responses to personality measures. Instability threatened the psychometric properties of personality measures, such as reliability, and this motivated researchers to find the source of instability within the measure. Perhaps more importantly, instability in measurements threatened the general assumption of personality theories that traits were stable over time (Winter & Barenbaum, 1999). This is evident in an early paper on personality theories in which the author points out that whenever instability in a test is identified, "we tend to look for the explanation in terms of the inadequacy of the tests" (Hendrickson, 1934, p. 246). In another strikingly apt observation from this paper, the author insists that the measurement of personality is a function of the society or context it is within. "The social setting of the measurement must be regarded as part of the situation to

which the subject is responding", which could be read as a foreshadowing to later criticisms of personality theory which argued for the importance of the situation (e.g., Mischel, 1973).

Before diving into the literature, a few notes. Many of the initial personality measures, the response scales generally had a dichotomous "True/False" or "Yes/No" response format. In some cases, participants were able to leave the response blank. Furthermore, the personality tests that existed often consisted of adjective checklists.

The first studies which sought to demystify features of unstable items focused on aspects of the semantic content of the item. In one of the earliest studies, Bain (1931) analyzed data from undergraduates who completed 61 personnel information items at two time points separated by two and a half months. The author of this study focused on the nature of the information that was requested. There were three categories of items: factual family data, factual personnel data, and subjective personal data. Results indicated that factual personnel data were most stable across time points, which was hypothesized. Overall, Bain found that one-quarter of the responses had changed. Similar results were found by M. Smith (1933) in a study aiming to replicate Bain's results. Both of these studies included males and females in their samples and analyses and found that women tended to have more stable responses than men, though the explanations offered did not offer much insight. Interestingly, the two studies differed on which items were identified as being stable or unstable. Smith suggested that more careful study be conducted on items used in questionnaires in order to understand which properties were related to instability.

In another study examining stability in personality questionnaires, Lentz (1934, p. 351) observed that "the more [stable] items are more simple in meaning, more definitely stated, and...more in the realm of the students' interests". Lentz, however, did not conduct any formal statistical analyses which took these item features into account, and his descriptions were based on his impression of the items. The results from this study indicated that there was great variability in the *amount* of instability among subjects. In other words, some participants did not change any responses, whereas others changed nearly every response. The author noted that this finding suggested that the instability was "a function of the persons," while admitting that this was a hypothesis that would need further investigation (Lentz, 1934, p. 354).

Frank (1935) studied stability in a sample of males using the Bernreuter Personality Inventory (Bernreuter, 1935). Items were coded as positive, negative, and neutral based on "general agreement," or the similarity in responses (Frank, 1935, p. 322). Items which were answered "yes" by at least 75 percent of individuals were positive, items answered "no" by at least 75 percent of individuals were negative, and the remaining items were considered to be neutral. The results showed that neutral items were more likely to change from one point in time to another compared to positive or negative items. In another study, a sample of children completed four administrations of a personality inventory across six weeks. The authors compared items which were changed by the largest percentage of the sample to those

changed by the smallest percentage of the sample (Pinter & Forlano, 1938). They did not find any differences in the items in regards to how vague the item was, the frequency of the situation described, the difficulty of the item, or the length, features which the researchers presumably coded for themselves. They also did not find any differences in consistency in responding when grouping the children by personality traits. A similar study included an "analysis of the changed responses with reference to the nature of the item" (Hertzman & Gould, 1939, p. 346). The researchers used two dimensions with three levels. The first dimension was related to aspects of the item content domain: whether it described something factual, physical symptoms, or feelings/attitudes. The second dimension was frequency and coded items dichotomously based on whether they used frequency words, such as "often" or "ever," or did not specify a frequency. Results indicated that factual items were the most resistant to change and that items including the word "often" had the largest amount of change.

Glaser (1949) sought to determine whether instability was a problem unique to personality tests, or whether other types of tests (i.e., tests having different content) showed similar instability. In a comparison of a personality test, an intelligence test, and an interest test, no significant differences were found in regards to inconsistency of responses. These results could be interpreted as suggesting that the use of language in tests (as opposed to another type of stimulus, such as pictures) was related to instability, though this was not something the authors noted. Inconsistencies which were found to occur for each type of test could either be ascribed to attributes of the particular individuals taking the test, the formatting features shared by the three tests, or an interplay between these two factors. Mitra and Fiske (1956) looked at the stability of responses in an Interests Test and an adjective checklist within a sample of men in the military. While the authors did not call attention to the effect of item features, some findings in regards to item features can be derived from their results. Items which were negatively worded on the adjective checklist showed the strongest relationship between the shape of the distribution of responses for the item and the frequency of change. That is, items which were more uniformly distributed were more likely to change. However, items which were about interests relating to "daring" or "mechanical" showed weaker relationships between the distribution and the frequency of change, suggesting that there was something about these types of items that might have explained the instability better than the distribution.

In a study which attempted to uncover rules which lead to more stable responding, the authors first selected inconsistent and consistent items from two administrations (Owens, Glennon, & Albright, 1962). Then, they examined the items in what was described to be a more qualitative approach, and developed 12 rules they thought differentiated the consistent items from the inconsistent items. Then, they chose the four rules they felt were most revealing, and had graduate students rate the items in terms of these rules. The

rules included brevity, or length of the item stem, numbering of the responses, exhaustive response options, and neutrality.

R. B. Kuncel and Fiske (1974) studied instability in relation to characteristics of the individual participants, as well as features from the questionnaire items. The methodology used to code the items was based on reports from the participant on five features. "New elements" asked the participants to indicate if their interpretation of the item was different upon seeing the item a second time, thus causing their response to be changed as well. "Recent" asked the participant to indicate if an event which had taken place recently had influenced their choosing of a response which they would typically not choose had the event not occurred. "Meaning" asked if the item was ambiguous, while "Apply" asked about whether the individual had difficulty applying the item to their own experience. "Difficult" asked the individuals to indicate how difficult it was to answer the item. Finally, the participants were able to provide a question mark if they were not able to choose a response to the items, which were dichotomous. Results indicated that items which were rated as having New Elements or as Recent were less stable.

Bond (1987) sought to determine whether item properties were causes for instability. Using content scales developed by (N. Wiggins, 1966), this study included maladjustment items from the MMPI as well as items which were judged to be neutral with regards to maladjustment. For example, an item like "I enjoy detective or mystery stories" was considered to be neutral with respect to maladjustment. Also included in this study were ratings indicating their social desirability scale value. This study offered a contribution to the study of item features by coding the neutrality of an item in regards to a *specific* trait, which, the author argues, reveals more about the item content. The findings indicated that the maladjustment items had more extreme social desirability scale values, but that they also showed more inconsistent responding. When matching the two types of items on their endorsement rate, maladjustment items still showed inconsistent responding.

There appears to have been a sort of hiatus in the study of item features for some time within personality literature, for reasons that are not clear. However, Archer and Gordon (1994) reinvigorated this area in a study which modified items on the MMPI in regards to content and structure as a way to improve the stability of items. The results from this study however did not indicate that "improving" the items would result in more stable items.

While the current study does not look at unstable responding, it is important to get a sense of various applications for item features. Having a framework for characterizing the linguistic features of items would assist researchers interested in understanding stability. Such a framework could also help the theoretically-focused psychologists who are interested in testing and proposing cognitive theories for instability. A result of research on instability is the identification of ambiguity as being potentially related to unstable responding, or at least theorized as being a cause of unstable responding. The next section will discuss ambiguity and how it has been defined and studied.

Ambiguity

Ambiguity is an item feature related to response stability, but studied in enough detail to warrant a separate section. The ambiguity of test items was an early challenge in developing personality inventories (Allport & Odbert, 1936; Broen, 1960; Eisenberg, 1941). Allport and Odbert (1936, p. 449) describe two types of ambiguity: the first refers to differences in interpretation of items across individuals. For example, for the item "I am a shy person," one individual might take "shy person" to mean being quiet or introverted and necessarily mutually exclusive from being extroverted, while another individual might think of this as being more along the lines of feeling nervous about being around others, but not necessarily being quiet or introverted. The other type of ambiguity refers to intra-individual differences, which relates highly to instability. For example, considering the item "I am a shy person" again, an individual responding to this item might first focus on the word "shy" and whether or not they are ever shy and respond accordingly. However, upon seeing it again, they may focus more on the words "shy person," and respond about the extent to which they believe their shyness qualifies them as a shy person.

The interest in ambiguity is tied heavily to the use and interpretability of personality questionnaires. The intention of these questionnaires is for individuals who are similar on their "true scores" of a particular trait to provide relatively similar responses on the questionnaire. If individuals who are similar in their true scores have drastically different interpretations of the same items, these items would be of little value. Rotter (1954, p. 257) goes so far as to say that homogeneity of participants' interpretation is an assumption underlying the use of personality inventories. In terms of methodology, ambiguity has been operationalized using two (or three) different approaches. First, there are sample specific or data-driven approaches. These could be considered one approach since they both depend on having a sample, or they could be further divided into two approaches. Specifically, some research has used subjective ratings of the ambiguity of each item, and these ratings are completed either by the participants themselves, or by the researchers. Another data-driven approach uses statistics derived from the sample, such the percentage of individuals who changed their response across two administrations of the same questionnaire. The second broad approach is intrinsic, which can be done without collecting ratings or a sample of data. This approach uses various linguistic categorizations, such as the number of definitions of a word, to represent ambiguity.

However, since questionnaire items were written to be clear, researchers wondered to what extent was ambiguity a problem in personality assessments. To investigate ambiguity, researchers first sought empirical evidence for the existence of a variety of interpretations of questionnaire items. Eisenberg (1941) did just this using 25 binary items from the *Thurstone Neurotic Inventory*. Participants in this study were asked to describe the meaning of each item and their reasoning for their response. In the author's analysis, Eisenberg found

that there were at least two different interpretations for every item. He also found that two individuals sometimes had the same interpretation of an item and the same reasoning for their response, but different responses. This was termed "overlap" and occurred in 44 percent of responses. He also looked at equivocal responding, which he defined as situations where the respondent gave a "yes" or "no" response, but revealed in their reasoning that they did not meet all criteria described in the question. This occurred for 22 percent of responses. Finally, 13 percent of responses were non-responses. Overall, individuals who were considered to be "neurotics" showed little overlap compared to other groups. Eisenberg interpreted these results as indicating that a high score on the measure could be understood as an indication of neuroticism, but that lower scores could not be interpreted as an indication of "stability." This study showed that there was indeed ambiguity, which rendered the interpretation of any given score tenuous. In contrast, in a similar study, Benton (1935) did not find that individuals differed in their interpretations of items on a questionnaire, even when individuals differed with respect to the trait of interest.

In one of the most important papers from which the current investigation draws, Goldberg (1963, p. 470) posited ambiguity as a "crucial component of item stability," and developed a model for item ambiguity. Some researchers at this time had been using the variability of an item across time points as a proxy measure for intra-individual ambiguity. However, while it makes logical sense that more ambiguous items would lead to less stable responding, Goldberg argued that it is problematic to simply use the change in endorsement as a measure of ambiguity, as doing so could lead one to discard items which were unstable. Using the change in endorsement is a data-driven, or empirical metric for ambiguity which does not directly ask participants for their perceptions about the ambiguity of the items. Goldberg and other researchers such as Broen (1960) found that items which were endorsed by a large majority (i.e., 90-100 percent) were more stable. However, these items were not able to discriminate across individuals. Because these items consisted of extremes of the given attribute, they showed high endorsement in the population. A scale with perfect temporal reliability which did not discriminate individuals with respect to the attribute of interest would be of no utility in most cases. Both Broen and Goldberg pointed out that personality items which showed the greatest instability were also the most discriminating items. Goldberg cautioned against defining ambiguity based only on the data obtained and suggested that we also look at the item properties in order to maximize discriminating power and minimize varying interpretations within an individual (Goldberg, 1963).

Goldberg (1963) sought to improve the definition of ambiguity and created an index for ambiguity (Ambdex) that could be used by researchers to select items which are optimal for discrimination, but which are stable over time. The original development applied only to dichotomous items. To calculate Ambdex, the researcher must have at least two administrations of a particular item pool. From these two administrations, the researcher can calculate the endorsement frequency of each item, as well as the percentage of the sample

that changed their response to the item upon seeing it a second time. Goldberg goes on to derive this statistic using these calculations. Ambdex is essentially a measure of ambiguity which is based on the instability of the item, but corrects for the endorsement frequency. Goldberg argues that items which describe "some aspect of behavior that is directly observable" are more stable and less ambiguous than items which describe less observable states, like an attitude or feeling (Goldberg, 1963, p. 182).

Goldberg (1963) offered several postulates as a part of his ambiguity model, one of which was that "the greater the inter-individual variability in [the item], the greater will be the intra-individual variability" (Goldberg, 1963, p. 486). In addition, Baxter and Morris (1968) sought to test this proposition suggested by Goldberg (1963) that inter-individual ambiguity would be positively correlated with intra-individual ambiguity, as well as the assertion that inter-individual ambiguity would correlate positively with the ability of the items to discriminate while intra-individual ambiguity would correlate negatively. Both of these propositions were confirmed. J. S. Wiggins and Goldberg (1965) investigated 8 item properties and statistics which had previously been explored in the literature using the MMPI. These included group endorsement percentages, social desirability as measured by ratings, dispersion of social desirability ratings, stability across time, item ambiguity, direction of deviance, item serial position, and grammatical classifications. Of these characteristics, five are based on statistics which must be derived from the data. This paper uses the Ambdex statistic developed by (Goldberg, 1963), but calculates the statistic separately for men and women.

Expanding on his research on ambiguity, Goldberg (1968) examined additional indices of ambiguity using an adjective checklist. The indices included Ambdex, endorsement percentages, omission percentages (how often the item was left blank), stability percentages from retest, ambiguity ratings from the participants, and lexicographic information. Notably, of all of the operationalizations of ambiguity tested, lexicographic information was the only type which could be derived without having a sample of data. The lexicographic information represented the "richness" of the item's meaning and was calculated using information from the dictionary. Results indicated that the omission percentages were most highly correlated with individual's ratings of ambiguity. Conversely, the lexicographic indices tended to not agree with the ambiguity ratings provided by individuals. Certain adjectives had many potential meanings, but were not necessarily considered to be ambiguous. For example, "persevering" was considered to be lexicographically non-ambiguous, but appeared to be highly ambiguous for female responders (Goldberg, 1968, p. 289). Finally, Goldberg and Kilkowski (1985) examined a set of synonyms and a set of antonyms and found that antonym pairs had less consistent interpretations than synonyms. That is, antonym pairs tended to be more ambiguous.

Payne (1974) used four item features to represent various aspects of ambiguity. These included ratings completed by ten judges for global ambiguity, estimated stability, social

desirability, and reference to behavior. These ratings were gathered from six rating sessions, with an interval of two to five days in between each session. The ratings were then averaged for each item and each feature. This study also used the item length, whether it contained a frequency qualification (e.g., "sometimes","often"), and whether it referred to a recent past or more distant past as indices of ambiguity. The author considered these to be non-subject related features of ambiguity.

The research on item ambiguity has several takeaways. First, ambiguous items on a questionnaire do not automatically render the questionnaire obsolete, rather, certain types of ambiguity are less problematic than others. Exploration of ambiguity has also contributed various theories and methodological approaches (e.g., Ambdex). However, research on ambiguity has not yet specified the properties which make an item ambiguous, and tend to focus on whether a word has multiple meanings. Some researchers have operationalized ambiguity in terms of number of dictionary definitions, but this is a weak proxy at best for ambiguity. Dictionaries often contain archaic definitions. Moreover, the operational definition of ambiguity in this research does not clearly differentiate between ambiguity and vagueness, nor was structural ambiguity considered (for more discussion of ambiguity, see the section on Psycholinguistics). The lack of the consideration of structural ambiguity is most likely related to the use of adjective checklists. Future research in this area will need to further elaborate on the definition of ambiguity and differentiate ambiguity and vagueness and one focus of the current project is to consider the ways in which we might measure ambiguity without directly asking participants.

Response Styles

Just as item response stability inspired much of the research on itemmetrics, response sets or response styles have led researchers to a more careful consideration of item properties. Cronbach (1946, p. 476) introduced the theoretical definition of a response set as "any tendency causing a person consistently to give different responses to items than he would when the same content is presented in a different form". Response sets, response styles, and response bias are terms that all refer to the same general idea and will be used interchangeably in the present paper.

The idea of a response set has permeated the literature in personality and extended to other areas of psychology. In fact, there is continued debate as to whether response sets or response styles indeed affect responding. A number of response styles have been proposed, for example: extreme responding, mid-point or neutral responding, malingering or faking, and response bias with respect to an individual characteristic such as sex or culture. Two frequently discussed response styles are acquiescence and social desirability. Acquiescence is the tendency for participants to choose an agreeable response (i.e., "yes" or "agree") and thus have their true responses contaminated by their response bias. Social desirability response styles describe a tendency for an individual to provide responses in line with what they

believe to be socially desirable. In the context of a questionnaire about taboo behavior, an individual may under-report their behaviors or frequency of a particular behavior because that behavior (or the frequency) is not socially desirable. See Paulhus (1991) for a detailed review of research on response styles.

It is not completely clear from the research whether response styles are simply traits that exist within the individual, or whether they are evoked by the testing context. Based on the model we have proposed, we would speculate that response styles are a combination of individual characteristics that are evoked within the context of the testing situation. The content of the questionnaire, the form of the questionnaire items, the traits of the individual, and perhaps the culture they were brought up in are likely to interact to result in a pattern of responding that is caused by something other than the underlying construct of interest.

Response styles have particular relevance to the discussion of form and content because their identification led to various design features in questionnaires which intended to minimize or detect the influence of response styles. As discussed by Paulhus (1991), certain design features of questionnaires were developed with the express purpose of controlling for various response biases. Indeed, the research on response styles has drawn attention to the way in which items are written. We will limit our discussion to the two most well-known response biases: acquiescence and social desirability.

As an aside, the reader should be cognizant that the idea of a response style has been controversial since its inception. Rorer (1965) offers a review of the evidence against response styles representing any trait within the individual, and other researchers continue to critique the theory and the meaning of findings related to this line of investigation.

Acquiescence Acquiescence is understood to be a response style where an individual tends to agree with statements rather than disagree (e.g., Paulhus, 1991). Lentz (1938) originally noticed that some individuals had a tendency to respond "true" more often than "false" on a true/false questionnaire. He proposed this tendency as a trait called "acquiescence." Research on acquiescence has been fruitful in the personality literature since acquiescence could threaten the validity of the instrument being used (Cronbach, 1942). It has also been of interest from the perspective that acquiescence is related to some underlying individual difference.

One of the most pervasive innovations developed in order to detect or deter acquiescence is reverse coding or reverse keying of items. This design feature is not limited to personality assessment. Scales were initially designed so that responses were keyed in the same direction. For example, on a depressive symptomology questionnaire, a higher response for any item might mean higher depressive symptomology whereas a lower response for any item indicates lower depressive symptomology. In a test designed in this way, detecting acquiescent responding from severe depression would be impossible. Therefore, reverse-keying, or

reverse-coding of half the items allowed for the researcher to have a better idea of whether acquiescence was an issue.

Bentler et al. (1971) contrasted two types of acquiescence which were called agreement acquiescence and acceptance acquiescence. Agreement acquiescence leads an individual to agree with all types of items on a test, even when there are items which have negations. Acceptance acquiescence is when an individual endorses all qualities as accurate descriptors of themself. This occurs when an individual says that they are "happy" as well as "sad," but disagrees that they are "not happy" and "not sad" Paulhus (1991, p. 47). As Paulhus (1991) discusses, acquiescence in general is of greater concern when asking participants about their attitudes or behaviors (e.g., survey research) than on personality assessments. Paulhus argues that this is because the items on a survey about attitudes are more often looked at specifically whereas personality items are composited. That is, it seems that because survey research is often concerned with single item responses, acquiescent responding to those items would highly distort the results. Regardless, it is one of the most well-known response styles and researchers attempting to control for it in personality have discussed item features. For example, when trying to control for acceptance acquiescence, a researcher might consider including the positively worded item "sad" as well as the negation "not sad." However, the researcher cannot simply add a negation, instead, they must state the item as an assertion and find an antonym or conceptually opposite version of that trait, such as "happy" or "content." Adding a negation in some cases does not equate to the precise conceptual opposite of the trait. For example, "I am not sad" does not equate to being "happy". Someone who is "not sad" might be angry or anxious, which are very different from being happy. Sometimes can be quite difficult to come up with a conceptually opposite term for the trait of interest. Typically, preliminary studies are done to find an appropriate assertion. Schuman and Presser (1996) provide a more thorough exploration of this topic.

Other approaches to the consideration of acquiescence bias have included the development of a separate scale to measure a trait which caused the tendency to acquiesce. However, acquiescence is potentially domain-specific, and thus, none of the scales to date are considered to be of much use (Paulhus, 1991). In other instances, large assessment batteries have been designed in such a way that allows the responses to be checked for acquiescence across all items. Underlying this approach is the assumption that acquiescence is a trait within the individual, rather than something that exists in the item. If it is indeed the case that acquiescence is a trait within the individual, then a scale to measure this bias is extremely valuable outside of personality research. For example, when studying social attitudes, or any sort of evaluative survey research, acquiescence could be very problematic (e.g., Paulhus, 1991).

Generally, there is little research on acquiescence which has examined or proposed features in items which might lead to acquiescent responding. Instead, it has looked at ways to modify or write items so that acquiescence can be detected. Condon, Ferrando, and

Demestre (2006) examined linguistic and structural item characteristics to determine which types of features may be related to acquiescent responding. Their research was based on earlier hypotheses which suggested that longer and complex items tended to elicit acquiescent responding. The results provided evidence supporting that hypothesis. Bentler et al. (1971) also modified the wording of items as either positive or negative in order to determine whether there were different types of acquiescence. In a related investigation, Kulas and Stachowski (2013) looked at which characteristics lead individuals to choose middle category responses. Their results indicated that items which required additional contextualization for a response as well as items which were not clear were more related to middle-category responding than were characteristics related to traits. There is a need for more research which considers item characteristics as well as response category characteristics. One might imagine that a response scale which has the "agree" category closer to the item stem would be more likely to see a bias for acquiescence than one where the response categories are stacked or where "disagree" is closer to the item stem. If an individual is prone to acquiescing, they will see the "agree" category first, perhaps before they have even read the other options. If this were the case, one could theorize that factors such as fatigue, which could apply to any given participant, might lead a participant to tend to agree. Currently, it seems the theory is that acquiescence is a trait in and of itself that individuals have. Results that show complex items leading to more acquiescent responding can be seen as challenging this assumption.

In sum, the recognition of acquiescent responding has led some researchers to examine item characteristics as a source, but more often it has led to modification of item features in order to limit or detect acquiescent responding. There is a need for more research which considers item features which lead to acquiescent responding in order to better understand this phenomenon empirically. Equally important is the consideration of which types of questionnaires are most likely to result in acquiescent responding. To relate this back to our survey interaction model, we would hypothesize that a survey context where the participant may feel that their responses could be easily observed by others would be more likely to result in acquiescent responding.

Social Desirability Personality psychologists have often discussed social desirability as a response style which could affect the interpretation of results. The earliest exploration of social desirability was done in a study by Edwards (1953), where one group of participants provided ratings on the social desirability of certain items, and another group responded to those same items. The authors found a .871 correlation between the ratings and the endorsement of the item. One interpretation of these findings offered by the authors is that traits which were desirable might be more widespread in the population. They also suggest that individuals who take the test are, for some reason, driven to provide a more desirable perception of themselves. This raised the question as to whether this was a result of the

testing situation, or alternatively, a result of some trait that varied across individuals. It appears in the literature that social desirability is usually interpreted as existing within the individual, influencing their responses. Some researchers have developed scales to measure social desirability as a trait. For example, the Balanced Inventory of Desirable Responding (BIDR) is a 40-item measure that measures the factors of self-deceptive positivity and impression management (Paulhus, 1984).

However, some researchers argue that social desirability is something that exists within the item. That is, item content may describe activities or traits which are simply less desirable for the majority of individuals and thus lead to certain response patterns. The first hypothesis is more thoroughly studied, though either possibility is concerned with item properties in some way. Minimizing any type of response bias is an important goal for researchers and thus, understanding the characteristics of items that lead to responding in a socially desirable way has colored some of the literature on this topic.

Edwards (1957) discusses social desirability within personality research at length. Edwards offers a "Social Desirability Hypothesis," in which he argues that the tendency for participants to respond in a socially desirable way is more important in personality research than acquiescence or the tendency to disagree. This is because Edwards recognized that the social desirability of an item depended on the response. For some traits, the more socially desirable response was just as likely to be "true" as it was to be "false" (Edwards, 1957, p. 108).

The conceptualization of social desirability as situated within the individual relies on there first being an understanding of what a socially desirable response *is*. Edwards (1953) suggested that the relationship between socially desirable responses and endorsement in the population is monotone increasing. That is, as the social desirability of a response increases, so does the endorsement in the population. Using dichotomous or trichotomous response categories, as was common in personality questionnaires, may have reduced some of the complexity in defining this issue.

An example of the complexity of defining social desirability might include an item which mentions a taboo or illegal behavior. For example, imagine a question about drinking behaviors which asks someone how often they drink on a 4 point scale from "none" to "every night." In a college population, the social desirability of responding "none" might indeed be lower than "every night," but the endorsement of "none" might be higher than one would expect based on social desirability values if the respondents are under the drinking age. In the general adult population, these responses may have approximately equal social desirability values. In a population where the cultural context forbids drinking, the social desirability values are likely to look very different than in a cultural context where drinking is encouraged.

Messick (1960) conducted a factor analysis of social desirability to determine if it was indeed a single dimension, and found that there were 9 factors. The authors interpreted the

factors as representing "different points of view as to what is desirable" [p. 287]. N. Wiggins (1966) extended this work and found that there were 6 factors, or "idealized individuals," suggesting that there was not a single dimension of social desirability.

Jackson and Messick (1958) argue the importance of the consideration of response styles in addition to the content of personality questionnaire items. The authors mention acquiescence, over-generalization, and a tendency to respond in a socially desirable way, as types of response styles. For example, Stricker (1963) describes social desirability as a response style in which the respondent chooses responses which conform to certain social norms. On the other hand, Messick and Jackson (1961) investigated social desirability ratings of college students for MMPI items to determine the range of perceived social desirability. The goal in this study was to determine how these social desirability ratings may be useful for scale development. The authors suggested that their findings could be used in order to measure socially desirable response styles. Other researchers have described social desirability as a feature of the item itself(e.g. J. S. Wiggins, 1962; J. S. Wiggins & Goldberg, 1965).

The conceptualization of social desirability as an inherent item feature vs. the conceptualization of social desirability being a response style reveals two methodologies, respectively. Generally, one methodology uses a group of "judges" who provide ratings for the item as either socially desirable, neutral, or undesirable. The average is then taken as the degree of social desirability of that item. The use of this method is related to the tendency of many of the personality measures to have dichotomous response options which included either endorsing the trait or not endorsing the trait. This situation makes it easier to think of an item in terms of social desirability of the trait itself, rather than thinking of the social desirability of a particular response. Moreover, as noted by N. Kuncel and Tellegen (2009), there is an assumption that the rating for responding "true" allows one to imply the social desirability of responding "false" to the same item. In other words, this approach is problematic because it only considers the social desirability of full endorsement of a particular trait, rather than more graded responses. Another method includes correlating responses from an item with a separate measure of social desirability.

In a more recent study which challenges traditional methods of measuring social desirability, N. Kuncel and Tellegen (2009) define socially desirable responding as "Behaving in a manner that is consistent with what is perceived as desired by salient others. This group of salient others is likely to shift from situation to situation (work, school, marriage), although many behaviors are considered desirable in nearly all situations"(p. 202). In addition, Kulas and Stachowski (2012) found results that suggest a more universal type of social desirability, that is, one that does not vary across people as a trait. Rather, they see the social desirability as existing somewhere outside the individual, perhaps as part of the broader cultural context. They argue for a cognitive model of the personality assessment process to better tease apart the source of social desirability.

Clearly, the discussion of social desirability is complex and is heavily dependent on theory from social psychology and personality psychology. A central question remains: is social desirability within the individual as a trait (or compilation of traits) which involves the prioritization of socially desirable responding, or can a particular type of item elicit socially desirable responding for any given individual, regardless of individual differences? Moreover, does an individual actually believe that a behavior is more or less socially desirable, or do they believe that *others* believe a behavior is more or less socially desirable? These are intriguing questions. In the view of our survey interaction model, we might conceptualize social desirability as rooted within the overarching cultural context, that is, what is and is not socially desirable is first and foremost determined by the local or broad culture. Then, perhaps the embedded context of the questionnaire may tap into that overarching cultural context more or less depending on the 1) topic of the questionnaire and 2) the susceptibility of that individual to be influenced by social desirability.

Summary

Within the area of personality psychology, there is a small literature which has sought to define characteristics of items and their effects. This research has made general contributions that may guide the current study. First, there is clear rationale and applications for studying item properties, e.g., stability and response styles. In addition, as summarized in Angleitner, John, and Löhr (1986), itemmetric researchers have used different strategies to identify and categorize item features. Angleitner et al. (1986) used four categories based on a theoretical model, rather than methodological: (a) logical relation between construct and indicator, (b) surface elements of the item text, (c) semantic item properties, and (d) aggregate item response parameters (p. 62). Other categorizations proposed were based on the methodology used. Jones and Goldberg (1967), for example, used two categories: (a) sample specific features were those derived from the participants' ratings and (b) intrinsic features were related to structural aspects of the item. Goldberg's classification was (a) item ratings, (b) lexicographic indices, and (c) test-retest statistics, and similarly, Payne's was (a) judgmental, (b) structural, and (c) metric (Goldberg, 1968; Payne, 1974). These groupings are determined by the methodology used to score them. Endorsement rates, or change in endorsement rates is one method to define features which is data-dependent. Features that are comprised of participants' interpretations and/or ratings of items with respect to various dimensions (e.g., social desirability) are also data-dependent methods. Methodology where the researchers give their somewhat subjective ratings of items are not data-driven. These ratings can be done *a priori*, however, they are subject to a multitude of biases. They also do not allow researchers to have a better understanding of structural features which may affect responding.

The personality research on item features has many limitations, some of which can be improved upon in the present study. One glaring shortcoming is that many of the study

findings or approaches could not easily be generalized because the measures used had a restricted format. Adjective checklists, for example, frequently occurred in the literature. Adjectives, as opposed to statements, have less room for variability and completely control for syntactic structure. Use of these types of scales therefore limited the features that these researchers attempted to investigate. An adjective, or even a symptom checklist, is not necessarily conducive for all types of self-report tests, even within one domain. Furthermore, adjective checklists are keyed in a particular way such that the respondent either completely endorses the item or they do not. No room is left for gradation. Thus, the discussion of social desirability within the context of adjective checklists assumes that the socially undesirable response is endorsement of the trait, or, in other words, the adjective itself is socially undesirable. How then does the researcher using a scale with five response options determine which responses are socially desirable?

Furthermore, when researchers used data-driven methodology such as ratings provided by the participants, there is a problem of circularity. If it were the case that that linguistic features *do* in fact affect responding, it follows that linguistic features would affect the ratings given as well. Of course, the use of ratings was typically limited to semantic or pragmatic features, such as social desirability, ambiguity, etc. These ratings, however, cannot give insight about the more piecemeal characteristics of the items that cause them to be read as ambiguous. We argue that the only way to get closer to studying the specific features of items which constitute a property such as "ambiguity" is by using objective, *a priori* item analytic or classification techniques.

Another shortcoming is that the purpose in studying item features was perhaps limited by the motivation to justify personality theory and personality measurements being valid. It is reasonable to connect item features to validity, as Goldberg and Wiggins did when discussing itemmetrics, however, it appears that this goal limited the types of features which were included. For example, the large focus on item stability and ambiguity can be framed as an attempt to provide evidence for the state-trait personality theories (Winter & Barenbaum, 1999).

A more mechanical critique might mention that the definitions of the features are not consistent and lack cohesion across studies. For example, ambiguity needs to be differentiated from vagueness and the different types of ambiguity (e.g., lexical, syntactic) should be fleshed out. Instead, the literature describes a multitude of senses and measures for ambiguity, but many of these operationalizations have not persisted. Our current study will attempt to address this disorganization.

A more egregious shortcoming is that personality research inconsistently considered demographic variables such as sex or the linguistic background of the individual. Instead, these individual characteristics were likely controlled for by limiting the sample through certain exclusion criteria. It may have been the *zeitgeist* which allowed researchers to turn a blind eye to non-native English speakers in particular.

The present study will build on the foundation established within personality research by examining finer-grained linguistic variables. These variables are primarily descriptive, and are not determined by rating items. However, we will attempt to operationalize certain types of features which may have originally been rated by participants and instead define them under linguistic variables. Some ratings completed by the researchers will be used, and there will be some inferences that are made by the primary researcher about how an item is perceived. The goal is to create a framework that can be used for any self-report measure which is linguistically based.

We can only speculate as to the reasons why research on the interaction between item features and individual characteristics has not been further integrated within personality. Nor is it clear why this line of thought has not appeared in other areas of psychology. Perhaps, as is often the case, alternative research questions were prioritized. Only recently have journals which cover the overlap between psychology and linguistics emerged. Having an appropriate journal to publish such research may have deterred other scholars from pursuing this area. Another possibility is the knowledge required to do such work. Perhaps having expertise in linguistics as well as psychology was a rare credential to come across. In any event, if itemmetrics is to be realized as a unique line of research in personality psychology, or in psychology more broadly, there is much work to be done.

2.2.2 Education

Education offers a different motivation and paradigm for examining item features. In general, there are three types of tests in education: ability, achievement, and aptitude (Stemler & Sternberg, 2013). We will largely focus on achievement and aptitude. Achievement tests intend to measure the mastery of certain educational skills, while aptitude tests intend to measure the potential for future performance in a particular setting (Stemler & Sternberg, 2013; Rodriguez & Haladyna, 2013). Both types of tests include verbal and quantitative items. Perhaps the most central difference between educational tests and tests in psychology is that educational tests have a incorrect response. In this sense, they are quite dissimilar from self-report measures. Issues such as response styles are less of a concern in the field of education, save the response style that could indicate cheating behavior. On the other hand, issues of test fairness are a large motivating factor for studying item features, as is the determination of item difficulty. Both of these motivations place differences seen across various demographic groups as an important consideration.

In 2001, the U.S. government passed the No Child Left Behind Act which increased the amount of standardized achievement testing in schools (Congress, 2001). This statute has raised the stakes in educational testing, as it outlines punitive actions against educators and schools if students do not perform at a certain level. Students are most impacted by this statute as they are evaluated and compared based on these tests at a national and, increasingly, international level.

Since the same tests are sometimes used across single states, and in some cases (e.g., the SAT) the entire U.S., the fairness of a single test across various demographic and linguistic groups has been called into question (e.g., McNamara, 2013). Thus, since 2001, there has been a great deal of emphasis in the literature on fairness of testing especially in regards to race and ethnicity and/or linguistic background of the individual (Eignor, 2013). For example, in Zieky (2013) wrote a chapter in the *APA Handbook of Testing and Assessment in Psychology* devoted to discussing fairness and equity in test development and Camilli, Briggs, Sloane, and Chiu (2013) wrote a similar chapter. As noted by (Camilli et al., 2013), the *processes* of investigating fairness and validity look similar, but fairness and validity are "distinctly different properties of test score interpretation" (Camilli et al., 2013, p. 571). The assessment of invariance, or the lack of differential item functioning (DIF), across sex and race is a testament to the focus on fairness in testing. However, researchers assessing DIF are frequently not interested in features within the *item* that may be responsible for differential functioning, rather, they simply want to improve their test (Elosua & Lopez-Jauregui, 2007). As explained by Camilli et al. (2013) "DIF is often found, but it is rarely interpretable. This limitation is well understood, but DIF techniques are generally considered to be useful as one step within a process of establishing a fairness argument" (p. 573). This is generally because whatever might be *causing* the DIF is likely to be unrelated to the construct being measured.

Still, there are other issues that arise that may lead researchers in education to be interested in the specific item features. Briefly, those in the field of educational testing have considered item properties in a variety of ways. For example, in mathematics testing, researchers have compared items which are symbolic to word problems. Other inquiries focused on the type of content, that is, whether there are pictures or text in an item. Researchers in education were often interested in how different item stem and item response structures affected the difficulty of items and the ability for that item to discriminate among students (e.g., Ace & Davis, 1973; Benson & Crocker, 1979; Hughes & Trimble, 1965; Dudycha & Carpenter, 1973; Violato & Harasym, 1987; Birenbaum, Tatsuoka, & Gutvirtz, 1992). Linguistically-based features which have been examined in regards to difficulty and discrimination in the education literature include item readability or reading level (often thought to be measuring linguistic complexity), whether the item stem was a complete sentence, and whether the item stem included a negation.

Because the current project is largely concerned with self-report measures, the discussion of the education literature will be truncated. The purpose of this section is to provide the rationale for the interest in linguistic features in educational testing and extend some of those arguments to the present investigation. Moreover, it may be possible to apply some of the methodologies used in educational research to the current project.

Readability

Determining whether a particular text passage would be appropriate for a certain set of readers remains an important task for instructors and test-makers in education. However, the time required to have individuals rate the reading level of the text was unreasonable, and having the instructor guess was not a desirable alternative. Thus, readability formulas were developed as a way to more easily calculate the readability of a text. As stated by Klare (1974), "a readability formula uses counts of language variables in a piece of writing in order to provide an index of probable difficulty for readers. It is a predictive device in the sense that no actual participation by readers is needed" (p. 64). Klare provides a thorough synopsis of the readability formulas that had been developed at this time. He goes on to describe the evolution of the types of the various formulas. For example, the Lorge formula was one of the first and was used to calculate readability up until grade 12. This formula was calculated by taking the average sentence length, the number of prepositional phrases per 100 words, and the "number of different hard words not on the Dale list of 769 hard words" (Klare, 1974, p. 67). This formula has since been revised several times.

Rudolf Flesch created several formulas for adult reading material. The grade placement formula used the average sentence length in words, the number of affixes, and the number of personal references to determine the grade placement (Flesch, 1948). He then revised the formula to include the number of syllables per 100 words, the average number of words per sentence, the number of personal words per 100 words, and the number of personal sentence per 100 sentences (Flesch, 1948). This reading ease formula was eventually revised again to include the number of syllables and the average number of words per sentence. It results in a score from 0-100, with higher lower scores indicating lower readability. The Flesch-Kincaid grade level formula is similar to the reading ease formula, but is rescaled to give a score in terms of U.S. grade levels (Kincaid, Fishburne, Rogers, & Chissom, 1975). However, this score does not have an upper bound. Also, this formula gives more weight to the length of sentences rather than the word length. This formula is the most well known algorithm available for calculating readability or comprehension of text. It has been used to determine the reading level of newspaper or magazine, or speeches made by politicians and is even built into commercially-produced word processors.

Beyond the Flesch-Kincaid, there are many other readability formulas which will not be discussed further here. There is a large literature on the shortcomings of these formulas and attempts to improve them. One challenge related to developing readability algorithms is balancing utility with validity. It is not clear *what* these formulas measure. Lenzner (2014, p. 678) uses comprehensibility, reading difficulty, and readability interchangeably "to refer to the effort required to understand the meaning of a text". Lenzner goes on to examine whether commonly used readability formulas perform well when using pairs of "problematic" item stems (e.g., vague, structurally complex) and improved versions of the same item. The

authors found that the readability formulas did not always agree, nor were they able to reflect the improvement from the revised items used in the study. These formulas claim to be measuring difficulty of reading, which, given the variables in the formulas, seems to most be related to the time it takes to read a sentence, followed by the cognitive load that comes with larger amounts of information. While fully exploring the construct of readability is beyond the scope of this paper, readability is a topic that is related to the present endeavor in that the goal is to describe text in terms of features (such as length, syllables, etc.) and then come up with some sort of composite, which is the readability score.

The most relevant shortcoming of the concept of readability is the lack of consideration of lexical features and finer structural features of sentences. Moreover, semantic and conceptual features could contribute to the overall complexity of a text, while not necessarily making the length of words or sentences longer. Formulas are not able to address any factors related to the interaction between the reader and the topic of the text (Rush, 1984). Or, in the case of Templeton, Cain, and Miller (1981), different texts could have varying levels of readability scores despite their difficulty, as rated by individuals, remaining constant. Bailin and Grafstein (2001, p. 298) argue from a linguistic perspective that readability formulas are not valid measures of readability, nor is readability a single construct.

Negation and Difficulty

Research in education has explored other types of features beyond readability for other purposes. One such purpose is determining which features contribute to an item's difficulty. Other purposes include identifying components of tests that contribute "construct irrelevant variance," as these sources of variance contaminate test scores and may render the test useless (McNamara, 2013). Negation, for example, is a feature which has received some focus in the literature.

Harasym, Price, Brant, Violato, and Lorscheider (1992) looked at negation within testing and whether negative wording had an effect on the item's difficulty and/or discrimination and found evidence that negation in stems as well as responses was problematic. Gronlund and Linn (1990) recommended that negations be used sparingly when writing test item stems or choices. Some studies have shown that these types of items had a longer response time and that respondents made more errors when the response format was true or false (e.g., Wason, 1961; Violato & Marini, 1989; Zern, 1967; Huttenlocher, 1962).

A particularly fruitful area for the examination of item features and how they bear on difficulty has been in computer adaptive testing (CAT). CAT requires a large item pool in order to ensure the test is secure (Enright, Morley, & Sheehan, 2002). Thus, attempts have been made to make the generation of new items more efficient. One possibility created by Haladyna and Shindoll (1989) is using "item shells," which are essentially a grammatical or syntactic structure that allows for the test maker to easily create items. However, researchers did not only want to generate items quickly, but they wanted to be able to specify the

difficulty of the items they were creating *a priori*. Another approach uses knowledge from cognitive theories to propose certain features that are likely to make an item more or less difficult (Bejar, 1993). A more fine-grained approach in the literature would examine a large number of features in order to understand which features are most related to difficulty.

As mentioned, research from cognitive psychology has contributed greatly to educational measurement research. A relevant stream of this research involves using cognitive theories to generate a large item pool. Embretson (1998) lays out a framework called "cognitive design systems" which would provide a way for test-makers to use cognitive theory when designing a test. To test this framework, Embretson generated abstract reasoning items and varied certain features. While this study did not use verbal items, the concept and evidence put forth is important to consider. Essentially, Embretson found that the tests developed according to empirically-supported cognitive theories were more valid than tests that were not. This finding is important since it adds more strength to the argument that item properties should be considered *a priori* as opposed to after the data come in. Additionally, it demonstrates the concerns over test validation as related to the test content (Embretson, 1998).

Lane (1991) designed a study to examine difficulty of different types of algebra problems based on cognitive theories. The items were word problems which differed in regards to complexity. One of the variables manipulated for the word problems was whether the story context was familiar and this variable was found to have an effect on how difficult the item was. Sebrechts, Enright, Bennet, and Martin (1996) explored "the varied factors that contribute to problem difficulty and to...performance [as a] way of evaluating the construct validity of measure of quantitative ability"[p 286]. This study recognized that there was an "interactive nature" between the problems on an exam and the individual's strategies. The design involved describing algebra problems in terms of their "attributes" which included such descriptions as distance rate and time problem, compound interest problem, graduated rate problem, or work problem (Sebrechts et al., 1996, p. 296). Findings indicated that some of the problem attributes had effects on the strategies students used to respond, as well as the errors. Interestingly, linguistic attributes did not appear to have an effect on the performance. It is worth noting, however, that these were GRE problems and the sample was fairly small. In a similar study, Enright et al. (2002) sought to determine whether changing the way that certain types of quantitative problems were written for the GRE would change their difficulty. Their results indicated that they were able to systematically manipulate difficulty of the items by changing certain features of the items, such as how many steps were required or the context of the problem described.

Verbal ability testing research has taken a similar approach. By looking at variables that are thought to be related to reading comprehension, or other verbal abilities, researchers hope to be able to generate items. Some features that were manipulated in this area of research include the text features such as the propositional density and the level of

vocabulary (Sonnleitner, 2008; Gorin, 2006; Embretson, 1999). The propositional density is the proportion of the number of propositions to the total length of the passage (Gorin, 2006, p. 397). As such, these features are often manipulated in the text passage rather than the actual item stem.

Summary

Research from education provides a few takeaways. First, issues related to defining readability and using readability formulas demonstrate the dangers associated with superficial analysis of sentences. However, interest in readability has motivated researchers to consider the different factors that may influence the level of reading ability required to understand a particular text. Like the current study, education research has utilized models from cognitive psychology to inform their methodological approaches. Moreover, the evidence that manipulating various linguistic features of the items on educational tests can indeed affect the difficulty is further rationale for studying item features in a different context.

There are, however, some disparate differences between educational research and research on mood and well being. To start, educational tests are not self-report measures. The purpose of educational tests is radically different from that of psychological self-report questionnaires. For example, we are not typically concerned with making a questionnaire item more or less difficult in the same sense that educational test makers are. Moreover, some educational assessment researchers, such as those affiliated with Educational Testing Services, have the advantage of access to large, representative samples of individuals, as well as the ability to implement experimental sections on standardized tests like the GRE. These researchers are therefore able to implement complex experimental manipulations without the typical logistical barriers of recruitment. Furthermore, their motivation to generate test items arises from the need to deter cheating, which is not as large an issue within mood and well being research.

Expanding upon the issue of experimental design being more common in educational literature, the manipulation of certain features is a systematic approach with the important advantage of stronger inferences. Again, given the statistical power that standardized testing programs have, and the fact that standardized tests often have entire sections dedicated to experimental items, there are few disadvantages. However, the experimental manipulation of psychological instruments is problematic. Using an experimental approach whereby the researchers purposely manipulate questionnaire items might certainly allow the researcher to make stronger inferences about which features impact responding. However, we argue that a primarily experimental or empirical approach is problematic when using a tool that has already been established and has been used across a variety of settings. Generation of new items is standard in educational testing, but to date, is not typically advantageous for self-report items. It is unclear how results from an experimental manipulation of item features would generalize.

2.2.3 Research on Item Features in Self-Report Questionnaires

Miscellaneous item features

Content related features, or semantic features, are difficult to fully separate from features which are structural or related to form. A feature such as item length can be labeled as a linguistic or structural feature of the item rather than content. On the other hand, the presence or absence of a frequency word (e.g., "always" "never" "often") could be considered as a linguistic feature since the underlying structure is affected, or it could be considered as a semantic or content related feature because the meaning is affected. To avoid these conceptual challenges from distracting, the current section will simply discuss other types of features which have popped up in the personality literature. Perhaps one way to differentiate these types of features is to look at the methodology. Structural/syntactic features can be coded without collecting additional ratings since they do not require interpretation. Content related or semantic features may be coded without additional ratings, or not. Moreover, whether the ratings are completed by the researchers themselves or by participants in a larger study can provide some indication to the presumed subjectivity of a feature. Attempts to distinguish features which are more related to content, from features that are more related to linguistic structure will be made.

Werner and Pervin (1986) examined features such as the area of functioning described in the item, the situation described in the item, the frequency of the behavior described in the item, and the time frame included in the item. The aim of the study was to determine the variety of items used in commonly used personality measures. The ratings were determined by three judges, suggesting that these features are more related to semantics, but are not considered by the researchers as being so subjective that ratings need to be collected at a larger scale. The results showed that for all the categories other than time, there was a variety of item types within questionnaires, and across questionnaires.

Payne (1974) included item content features such as the recency of the item, whether it referred to a behavior, and the frequency of that was described in the item. This study found that ambiguity was related the lack of a mention of behavior. However, there were no other significant relationships found between content-related features. The author concluded that these types of features would be easily manipulated and should thus be studied in the future. For social desirability, estimated stability of responses, global ambiguity, and reference to behavior, ratings from ten graduate students, taken at two time points, were averaged and used in the analysis.

J. S. Wiggins and Goldberg (1965) included a large variety of linguistic and structural item features, including sentence length, active vs. passive voice, tense, and person. The methodology used to characterize items, again, differed depending on the type of feature. Wiggins separated grammatical classification from the other categories of features, which included endorsement percentages, social desirability, stability, ambiguity, and deviance.

The grammatical classifications were not based on ratings. Actual data from various samples were used to determine endorsement percentages, stability, ambiguity, and deviance, and ratings from students were used for the social desirability variable. They found that sentence length, sentence structure, negation, and temporal frequency showed certain patterns and relations to other characteristics, but did not have any clear conclusions from these relations. The authors concluded that the grammatical characteristics used were not specific and that more refined procedures would be better able to study the effects of item characteristics.

Holden, Fekken, and Jackson (1985) examined a variety of features in personality measures that were related to item quality, and examined which structural and linguistic features might be related to quality. Quality was defined as item "goodness" and was operationalized using item criterion validity, content saturation (e.g., item-total correlations or item factor loadings), and item stability. These features were thus data-derived; linguistic features that were explored included negative wording, absolutes, self-referents, length of item, response latency, item desirability, item disguise, and response strategies. Results indicated that negations, absolutes, neutrality, and disguising were all related to lower item quality. Results also indicated that items which were very specific had less overall quality.

In addition to the features mentioned previously, Payne (1974) also included the length of the item to determine whether the length was related to stability. The author interpreted the findings to suggest that item length was not an important feature in the stability of an item. Goldberg and Kilkowski (1985) conducted an experiment where participants were given sets of synonyms and antonyms. One group received definitions of the adjectives. Their findings indicated that including the definitions did indeed increase the consistency of responding, suggesting that in some cases, additional words or explanations are helpful for participants to understand the intended meaning of items on a test. In contrast, Hamby and Ickes (2015) conducted a meta-analysis of seven personality scales and examined their reliability with respect to the average number of words per item and whether adjectives or statements were used. This study concluded that having fewer words per item as well as using adjectives resulted in better internal reliability. Notably, these authors conducted the analysis of item length using only items which were statements and found the same effect.

Reverse-wording, Reverse-Coding, or Negative Wording

On a self-report questionnaire measuring a particular construct, items are typically written so that the responses are keyed in the same direction with respect to the construct of interest. That is, if the questionnaire is asking about depressive symptomology on a dichotomous "agree/disagree" response scale, a selection of "agree" will be interpreted as endorsing depressive symptomology – and this will remain consistent for all items. Reverse-wording or reverse-coding refers to items on a test which are worded in such a way that the responses must be interpreted in the opposite direction from the rest of the items. Using the example of a depressive symptomology questionnaire, a reverse-coded item might ask something like

"I feel happy," where a response of "agree" would not be interpreted as endorsing depressive symptomology.

The use of reverse coding is intended to deter or detect acquiescent responding. Theoretically, having reverse-coded items would allow for the interpretation of the responses as being caused by the trait of interest rather than by a responding bias. Moreover, these types of items are also used as a way to have participants pay closer attention to the questionnaire. However, these items have caused a host of theoretical and methodological issues for researchers. In fact, reverse-coding may be one of the most frequently studied item properties in the literature. Here, we select a few studies which we believe to be illustrative of the literature.

Terborg and Peters (1974) wrote an attitudes questionnaire with two versions, one with favorable stems and the other with oppositely worded unfavorable stems. Participants had to respond to both versions, in counter balanced orders. The purpose of the study was to examine acquiescence and results indicated that the wording did have an effect on responses, but that there was no evidence for an acquiescence bias. Here, favorable and unfavorable wording was determined by the researchers, who reportedly made an attempt to make the reversal as logical as possible. One shortcoming of this study is the lack of examples of the two versions of item stems. This study is also of interest because the methodological approach was experimental, allowing for more control and for better inferences to be made about the results.

Chang (1995) conducted a study of negatively coded items on an attitude questionnaire with Likert-style responses. The authors first gave participants 8 items, 4 of which were reverse coded. Then, one week later, the participants were given the same items, but coded in the opposite direction. Results suggested that the wording had an effect on what the item was measuring, and the author suggests not using these types of items. Barnette (2000) examined survey items and manipulated the wording so that there were reverse coded items in three of the conditions. Another condition had a directly worded stem with a bidirectional response option. The results indicated much greater internal consistency in the condition with the bidirectional response option, and no negatively worded items.

Greenberger, Chen, Dmitrieva, and Farruggia (2003) examined the Rosenberg Self-Esteem Scale, which purports to measure positive and negative self image, which are each considered to be two dimensions. The authors in this study reworded the 10 items to be either all positively-worded, or all negatively-worded, and included those versions as well as the original version in their study. When completing a factor analysis, their results indicated a two factor model for the original version, and one-factor models for either of the revised versions. This suggested that the item wording on the Rosenberg Self-Esteem scale may result in an artifact – an additional factor. Roszkowski and Soven (2010) also found a separate factor for the negatively worded items and Solís Salazar (2015) found that including positive and negative items destroyed the internal consistency of the scales.

Ebesutani et al. (2012) used IRT to assess the utility of reverse-worded items on the Loneliness Questionnaire in children and adolescents. Their results were in line with other studies and suggested that reverse worded items had poor reliability, discrimination, and item information, relative to non-reverse-coded items.

In sum, reverse-coded items tend to cause issues for researchers. It is not altogether clear from each of these studies how the reverse-coded items were structured (e.g., do they contain a negation? are they logical opposites?), which presents an opportunity for investigating the features of reverse-coded items that may be causing the problems we have seen in self-report questionnaires.

Readability

In research on items used in psychological instruments, the use of readability formulas is advised against since items tend to be very short. Typically, readability formulas are computed on passages of at least 100 words, which is much longer than the typical item on self-report questionnaire (Flesch, 1948). The Homan-Hewett readability formula is an exception to this issue (Homan, Hewitt, & Linder, 1994). Hewitt and Homan (2004) Showed that this formula was effective for single items on a standardized test. The results from this study showed that across three grade levels, the readability of the item and the number of students who miss the item were positively correlated, suggesting that test items were assessing something other than content knowledge (Hewitt & Homan, 2004). Clearly, having a higher grade level of readability for a text can seriously endanger the validity of the test being used. Unfortunately, the Homan-Hewett formula has not been adapted into a computer applet allowing a researcher to input a sentence and receive a readability score.

Some research on self-report questionnaires has included readability. Jensen, Fabiano, Lopez-Williams, and Chacko (2006) examined the readability of 84 parent- and child/ adolescent report measures commonly used by clinicians. Their approach included removing repeated response sets and including the instructions as part of the text, and they used four readability formulas. Their results indicated a great amount of variability in reading level, as well as many measures having reading levels above the recommended 8th-grade level. McHugh and Behar (2009) examined the instructions and items for 105 self-report measures, and found that most measures were written above the mean reading level for the United States, that anxiety measures were less readable, and that the reading level for the instructions was high. Because some readability measures require the full sentences and some item stems are not written as complete sentences, the authors noted that their method involved rewriting the items using the anchor and the stem. They also noted that they did not include "numbered scales with anchored responses and other repeated response choices"(McHugh & Behar, 2009, p. 1107).

In another study McHugh, Rasmussen, and Otto (2011) looked specifically at evidence-based measures of anxiety for comprehension level. The authors used a program called "The

Question Understanding Aid" (QUAID) as well as readability scores and found that there were additional issues with these questionnaires that were not indicated by the readability formulas. Hamby and Ickes (2015) used a meta-analytic approach to examine readability of personality scales and found that readability tended to have no relationship with the reliability of the scale. Lenzner (2014) found that the readability formulas were not sensitive to the linguistic complexities that would likely affect the readability of the scale.

Summary

There has been some research on miscellaneous features within psychology, including structural features as well as semantic features. The most research has been done on whether an item is reverse coded. This area of research has shown that reverse-coded items can often be problematic, which offers some motivation for the present study. Finally, readability has been incorporated in psychological research, but there are drawbacks to using readability measures.

2.3 Individual characteristics

Thus far, the interaction between item features and characteristics such as sex and language background have not had much attention in the depression literature. However, there has been considerable research on individual characteristics and how those relate to depression. For example, differences in prevalence or symptom presentation across sex or linguistic background has received some attention.

2.3.1 Sex

Symptomology

It has been well established that depressive symptomology is more frequently reported among women than men (for a review, see (Piccinelli & Wilkinson, 1999)). Women also are more likely to report somatic symptoms related to depressive symptomology than are men. This appears to be a fairly robust finding cross-culturally. Some research has shown differential item functioning between men and women on items on questionnaires that discuss crying.

2.3.2 Language background

Testing

Some research has looked at differential item functioning with respect to linguistic background, or cultural background of the individual. In addition, the study of DIF in translated versions of tests has been of interest within the psychometric literature. However, while

there are ample empirical investigations of DIF across groups or across adapted tests, there is a scarcity of literature which attempts to identify the source or cause of the DIF (Elosua & Lopez-Jauregui, 2007, p. 40).

Symptomology

Several studies have indicated that lower levels of English Language fluency are related to higher reported negative symptomology. Rumbaut (1994) found that children of immigrant parents who were also less fluent in English reported higher levels of depressive symptomology on the CES-D. A number of other studies found this same relationship in different cultural and linguistic groups (e.g., Brown, Schale, & Nilson, 2010; Lee, Choi, & Matejkowski, 2013; Kim, Ehrich, & Ficorilli, 2012; Chung & Kagawa-Singer, 1993; Beiser & Hou, 2001; Hinton, Tiet, Tran, & Chesney, 1997; Khawaja, 2007). On the other hand, a few studies did not report a relationship (Huang & Spurgeon, 2006). Generally however, it seems that there is a relationship between linguistic background and the reporting of depressive symptomology.

2.4 Psycholinguistics

A final aspect deserving coverage in the current study concerns theories or models about how participants respond to questionnaires and process language. Briefly, important theories and concepts from psycholinguistics are discussed to give the reader some background on this topic.

Psycholinguistics is a relatively new field that is defined as the study of "how word meaning, sentence meaning, and discourse meaning are computed and represented in the mind. "[Psycholinguists] study how complex words and sentences are composed in speech and how they are broken down into their constituents in the acts of listening and reading" (O'Grady, Archibald, Aronoff, & Rees-Miller, 2010, p. 429). Research within psycholinguistics often focuses on response time for recognizing a particular word, or the effect of priming, and seeks to sort out different models of processing. In addition, the question of whether language processing is a unique component in the brain, or whether it is simply another type of cognitive process is an ongoing debate.

2.4.1 Universal characteristics of language

As summarized in Whitney (1998), there are six linguistic universals which are generally accepted. These will be briefly explained here. First, semanticity is the idea that language signals are symbols that communicate meaning. The symbols used in language have semantic content, they are associated with things in the world. Next, arbitrariness is the idea that a linguistic signal is arbitrary and does not have any resemblance to the thing which

the signal represents. This can be contrasted from a hypothetical linguistic system which might communicate using pictures which actually look like the thing they are representing. Discreteness is the concept that language signals are distinct and do not vary continuously. We do not say the word "dog" in varying pitches or volumes to indicate the size of the dog, instead, we use an additional discrete signal (e.g., "huge", "small"). Duality of patterning means that the patterns in language signals are occurring on two levels. One level contains the meaningful, discrete and arbitrary symbols, and the other level contains smaller units that do not have meaning on their own. In other words, our discrete words can be broken down into individual sounds that alone do not have meaning. Next, productivity describes the creativity inherent to human language. From the basic elements that make up words and elements, we are able to create many novel utterances. In fact, "most of what we say and comprehend consists of utterances that are original...[yet] comprehension is usually quite easy" (Whitney, 1998, p. 11). Finally, displacement refers to our ability to use language to communicate about things that may not be physically present, and we are also able to talk about events that happened previously, or that will happen in the future.

Semantic theories

The area within psycholinguistics that is concerned with the meaning of words and utterances relates to our present study. There are several theories about how meaning works. First, is the theory of truth-conditional semantics, originally developed within the field of philosophy (Montague, 2008). This theory is largely focused on the meanings of expressions rather than individual words. Two key components of this theory are *sense* and *reference*. The *reference* of an expression is the thing it stands for in the world, what the expression refers to. The *sense* of an expression is what results when combining the meanings of the individual words. A *proposition* is a relationship between two concepts that can be either true or false. Notably, this theory relies on a logical relationship between the expression and "possible worlds".

In opposition to this theory is the theory of conceptual semantics, developed by Jackendoff (1992). Jackendoff argues that meaning is a relationship between the symbols of a language and mental concepts. The meaning itself exists in the mind of the individual. This theory considers that the challenge for semantics is mapping between a syntactic structure and a conceptual structure. In terms of lexical semantics, conceptual structures are made from the concepts that we have stored about the world. Lexical decomposition is the breaking down of a concept into various features. For example, the concept we have about "dog" may consist of features like "animate," "barks," "mammal," etc. We know that a dog is an animal and that an animal is a thing, as opposed to a person, an action, an event, or place (Whitney, 1998). Part of this theory posits that there are some features that are so basic that they can allow for comparison across many concepts. These are what are called *semantic primitives*. A semantic primitive may be something like whether an object

is inanimate or animate. Another key part of this theory is that our knowledge of concepts includes images of objects as well as knowledge of how actions are performed.

An additional theory is the cognitive grammar theory that argues that mental models are the basis for meaning (e.g., Lakoff, 1987). One aspect of this theory is that there are basic-level concepts, similarly to the lexical decomposition in conceptual semantics. However, the concepts are organized at three levels: subordinate level (very specific), basic level (intermediate level), and superordinate level (more general). The basic level of this model is thought of as the most general level at which you can still form a mental image of the category. As an example, a "golden retriever" would be at the subordinate level, "dog" is at the basic level, and "mammal" is at the superordinate level. The other component of this model is the image schema, which is a model that can represent different types of interactions with objects in the world.

The current study will draw on these theories to explain the ways in which an item on a questionnaire is interpreted.

Ambiguity

Linguists have often discussed the issue of ambiguity in language (e.g., Whitney, 1998). There are several types of ambiguity, one of which is lexical or semantic ambiguity, which occurs when a particular word has more than one meaning. For example, "This building has many stories" is ambiguous because a "story" could refer to a level of the building, or it could refer to a myth or legend about the building, or it could even refer to physical books. There are other types of lexical ambiguity whereby the word has several senses which are closely related. For example, the verb "like" has many senses, all that have positive connotations and center around a similar meaning. Someone saying "How did you *like* your dinner?" is asking whether the individual enjoyed their dinner, and this is not quite the same as someone saying "What do you *like* to eat for dinner?" which is asking what their preferred meal is. And this is even slightly different than "What would you *like* to eat for dinner tonight?" All of these are slightly different from saying "I would like to eat something" and perhaps drastically different from "I really like her," which is more akin to "admire." A word is considered to be *polysemous* when it has multiple meanings.

Structural or syntactic ambiguity occurs when the sentence could potentially have two distinct structures, typically changing the meaning of the sentence. For example, the sentence "I saw that girl in my lab today" could have a syntactic structure where "in my lab" is meant to describe the "girl," and the underlying structure might look like "I saw that girl (who is) in my lab today". Alternatively, "in my lab" could refer to the place where where you saw "that girl".

Research in psycholinguistics has focused on the various types of ambiguity and how humans are able to resolve ambiguity. In fact, it is this aspect of language that has resulted in the most divisiveness within sentence processing models (e.g., Whitney, 1998). The

effects of ambiguity are theorized to be determined by two factors: meaning dominance and strength of context. Meaning dominance is the relative frequency of the various meanings that an ambiguous word has (Simpson, 1981). This may be a reason that earlier studies which defined ambiguity in terms of number of definitions did not find a great deal of evidence for this affecting stability. The strength of context is the degree to which the greater context of the sentence supports one meaning over the other. There is quite a bit of discussion over modularity and the way in which sentence processing occurs which is beyond the scope of this project. However, the evidence found so far does seem to suggest that sentence comprehension is sensitive to context. Thus, lexical ambiguity in a sentence is sometimes resolved using the greater context of the sentence, other times it is simply resolved based on what is the most common meaning of the ambiguous word.

Ambiguity is of importance in the current study for several reasons. First, given the focus on ambiguity in previous research, particularly in personality, continuing to investigate this property will allow for continuity. Additionally, we are interested in building a case for ways in which ambiguity might be operationalized that do not require direct feedback from participants, or data collected at multiple timepoints. Further, the property of ambiguity and how it might affect responses, as well as how it might interact with an individual's language background, is fascinating in its own rite. We will use the theories and research on ambiguity from psycholinguistics to inform our own definitions of ambiguity. Additionally, since ambiguity has been studied as a way to probe the way in which we process sentences, we have additional justification for including this as a variable which could play a major role in our survey interaction model. In considering these two types of ambiguity, we would expect lexical ambiguity to occur more frequently on self-report questionnaires than structural ambiguity. This is because most words that convey meaning (i.e., nouns, adjectives, adverbs, verbs) have at least two possible meanings, so lexical ambiguity is very common in any context. We expect that structural ambiguity would be detected when generating or reviewing the items.

Natural Language Processing

Briefly, one final area of research that is relevant to the present study is natural language processing. While technically an area of computer science, there is much overlap and borrowing of ideas and approaches from psycholinguistics that we have included it here. Natural language processing is a field concerned with human-computer interaction. The end goal is to enable computers to understand input from humans and generate language as well. This field has only existed since 1983, but many tools and approaches have been developed by researchers in this field for dealing with language.

2.5 Summary and Connection to Present Investigation

Generally, there are some themes that can be observed from the literature. The concept of ambiguity within psychological literature is fuzzy, and as such, whether the various operationalizations of ambiguity have any effects on responding within native English speakers is a question that has not been answered, let alone for non-native speakers. Next, psychologists have manipulated item properties with the intention to measure or reduce response biases, yet there appears to be some consensus that at least one of these features (i.e., reverse-coding) creates challenges for making claims that the test is unidimensional, and other unanticipated challenges. The concept of readability and the formulas designed to measure readability appear to be problematic as well, given the evidence to date from education and psychology. It is not clear what readability is measuring, and the findings from studies which included readability do not make a strong case for using readability as a metric, particularly at the item-level. Finally, a fairly large array of item properties have been suggested, and an even larger set of item features have been included in analyses. However, the literature seems to lack systematic definitions of these properties and there is little continuity across studies.

From a methodological standpoint, the literature offers multiple approaches for studying item features. These approaches can be separated into two main categories which we will define as sample-derived and structural features defined *a priori*. The sample-derived features depend on asking participants to rate the item stems in regards to various features and the structural features are based on using approaches from linguistics to describe items. The present study will elaborate upon the structural approach by defining additional features that could be considered. To date, there are no studies that examine the deep structure of questionnaire items, considered phrasal aspects of an item, labeled thematic roles and relations, as well as other features within the psychological literature. Nor are there studies that examine the ways in which these features interact with individual characteristics.

If we are to consider our survey interaction model of responding as well as the theories and evidence from psycholinguistics, the importance of the context cannot be understated. This creates some challenges for the researcher who intends to apply any findings about linguistic features across domains (for example, applying results from personality questionnaires to depressive symptomology questionnaires). Quickly, this assertion can place a great burden on the researcher studying item features because it limits the generalizability of their findings. However, the study of item properties is a research framework that is not young or new, yet, paradoxically perhaps, remains methodologically underdeveloped. The present study presents an attempt to further this research framework by enriching the methodology with knowledge from linguistics and psycholinguistics.

When selecting an instrument to use, we had several criteria. First, since our substantive research interests lie in heath and well-being as well as clinical psychology, we chose to use

a scale of depressive symptomology rather than a personality or ability assessment, about which we have less familiarity. The instrument had to have good psychometric properties and have evidence of utility in a number of settings, namely the general population as well as clinical settings. Next, an instrument which had item stems that were statements as opposed to adjectives would provide more variability in linguistic features. It was also desirable to have a scale that had at least 15 items, was not dichotomous, and that varied in their content as well as item features such as length. The inclusion of reverse-coded items was one feature we looked for in particular. We have chosen to use the Center for Epidemiological Studies- Depression instrument because it meets these criteria. The CES-D is widely used in research and practice, has 20 items, 4 subscales (including one that is reverse-coded), and has item-to-item variability. Of the most commonly used depressive symptomology tools, we could have instead used the PHQ-9, or the BDI-II rather than the CES-D. However, given that the purpose of the study was to examine item stems, the CES-D offers multiple advantages over these two other instruments. The Personal Health Questionnaire-9 (PHQ-9) is used in primary care settings as a screening tool, but only has 9 items, which are long and double-barreled The BDI-II has 21 items as well as linguistic variability, but the variability occurs within the response options rather than the item stems, which are adjectives.

A note on taking an experimental approach: while it is the case that one could control for a particular linguistic feature by manipulating the way an item is written, there are several theoretical and practical issues in taking this approach. To start, designing this type of experiment would require a researcher to have an inference about a particular feature. Examining reverse coded and non-reverse coded items is an instance where the researcher has an inference about the coding of the item. The design of such a study would require the researcher to create multiple versions of the same item that were equivalent in every way other than the feature of interest. This could be simple to construct in some cases. For example, to make the item stem "I was sad" reverse coded is quite simple, a negation is inserted to create "I was not sad". However, when modifying the meaning of the item, the syntactic structure has been modified, the length of the utterance has been modified, creating confounds. If there was enough evidence that certain syntactic structures did not have any effect on responding, or that the length of a sentence did not have any effect on responding, the researcher would not worry about these confounds. Furthermore, for any instrument already in use, modification of the items in a study would weaken the ability to generalize any findings.

2.6 Research Questions

2.6.1 What are the different ways to operationalize linguistic properties within the framework of psychometrics? How are these operationalizations correlated?

2.6.2 Are linguistic properties/features related to higher/lower endorsement of depressive symptomatology? Which ones?

2.6.3 Which properties are most related to higher/lower endorsement of depressive symptomatology?

2.6.4 Does language background and/or sex interact with certain item properties to lead to higher/lower endorsement of depressive symptomatology?

Chapter 3

Methods

This chapter will be broken into two parts. First, to address research question 1, the approaches used to code the CES-D will be described. Next, the approaches used to address research questions 2-4 will be described.

3.1 Focal Tool: Center for Epidemiological Studies-Depression (CES-D)

As described previously, we have chosen to use the Center for Epidemiological Studies Depression Scale to examine the effects of linguistic features and their interactions with individual characteristics/person effects. The CES-D is a 20-item measure of depressive symptomology (Radloff, 1977). The participant is asked to indicate how often they experienced various behaviors and feelings in the previous week. The item stems and response scale are provided in table A.1. The response scale is 4-points and has anchors which are "qualitative," (i.e., "Rarely"(0),"Some or little of the time"(1),"Occasionally"(2),"Most of the time"(3), respectively) as well as "quantitative" (i.e."<1 day"(0),"1-2 days"(1),"3-4 days"(2),"5-7 days"(3), respectively). A principle components analysis followed by normal varimax rotation determined four factors, interpreted as somatic symptoms, depressed affect, positive affect, and interpersonal problems. Radloff (1977) however recommended that the scale be composited despite evidence against unidimensionality because all of the items and the factors were related to depression. The author did address the administration of the CES-D in bilingual populations. As Radloff (1977) noted:

> There are some hints that understanding of the items may be a problem; there was a small but consistent correlation between the *CES-D* score and the interviewer ratings of understanding of the questions, independent of education of the respondant...Special caution is needed with bilingual respondents (Trieman, 1975). Further study of this issue is needed, with possible revision for simplicity of wording and removal of colloqual expressions (p.400).

Nonetheless, the CES-D has remained widely used in its original form for several decades. Shafer (2006) conducted a meta-analysis of studies using the CES-D and found evidence across studies for the four-factor structure. Additionally, there is evidence to support the compositing of the CES-D items (e.g., Wood, Taylor, & Joseph, 2010) and the composite or total scores are most often used in research (Santor, Gregus, & Welch, 2006; Shafer, 2006). Composite scores on the CES-D range from 0-60, with higher scores indicating higher levels of depressive symptomology

Analyses of the CES-D have indicated strong psychometric properties. The CES-D had high positive correlations with other scales designed to measure depressive symptomology, negative correlations with positive affect scales, and low correlations with scales designed to measure other constructs which one would expect to be unrelated to depressive symptomology (e.g., medications, social functioning, aggression). In addition, the CES-D correlated moderately with interviewer ratings of depression, but had low negative or zero correlations with interviewer ratings of cooperation and understanding of the question (Radloff, 1977). Three measures of internal consistency provided a lower bound of .85 in the general population sample and .90 in the patient sample (Radloff, 1977). In more recent study of a general population, Cronbach's alpha was .93 for the CES-D (Schalet, Cook, Choi, & Cella, 2014). For our sample, Cronbach's alpha was .88, which is similar to the internal consistency reported for a general population sample in the original study.

3.2 Linguistic Analysis

In this section, we describe the tools and approaches used to address the first research question:

2.6.1 What are the different ways to operationalize linguistic properties? How are these operationalizations correlated?

Using the exemplar *I felt hopeful about the future*, we describe each step of the linguistic analysis, as well as the correlational analysis. The full list of features in terms of how they were included as variables in the statistical analysis (i.e., dichotomous, continuous) appears in appendix A.

3.2.1 Tools Used for Analysis

Partially because of the recent growth of interest in Natural Language Processing (NLP), a number of tools have been developed which are freely available for researchers. These tools were used in the present study to verify our classifications in some cases and to provide linguistic features. The tools we used are briefly discussed.

Xerox Morphological Analysis Xerox has an open source tool which provides a full Morphological Analysis of input text (Xerox, 2016). We used this tool to verify our labeling of roots and affixes. This tool provides all of the possible parts of speech for the input, but does not indicate the correct part of speech (POS). It is up to the user, rather than the tool, to determine the correct part of speech.

Enju Deep Syntactic Parser We also used Enju, a deep syntactic parser which generates a phrase structure tree as well as the predicate argument structure (The Tsujii Laboratory, 2004). We primarily used this tool to verify our syntax trees and syntactic features.

FrameNet and WordNet In addition, we used the FrameNet lexical database developed by the International Computer Science Institute in Berkeley (Baker et al., 1998). This database was created based on the theory of Frame Semantics, which postulates that the meanings of words can be extracted by using a semantic frame, a description of a type of event or relation and the types of participants in it. Our interpretation of "frame" is similar to the concept of "sense." FrameNet describes a large number of frame types, and the other frame elements that can occur for that particular frame. For this study, we use the part of this database has annotated Lexical Units (LUs), or words, that are indexed in terms of the types of frames they can occur within. WordNet is a database developed by Princeton University for use in computational linguistics (Princeton University, 2010).

We used both of these tools since they seem to be based on two different semantic and/or syntactic approaches. For example, an alternative to X-bar theory in syntax is the idea that structure is determined by the lexicon. That is, the words themselves determine the structural elements, rendering something like a movement transformation as unnecessary. We refrain from delving into that particular debate here since this is not a theoretical linguistics paper, but one may refer to Bresnan (1978) or Fillmore (1968, 1971) for further reading.

Coh-Metrix Finally, the computer tool Coh-Metrix was developed to analyze text on over 200 measures related to cohesion, language, and readability, many of which had been previously established (Graesser, McNamara, Louwerse, & Cai, 2004). We used some of the measures directly from this tool in our analysis and we also used it to verify our own coding and labeling of the items. This tool is perhaps the most powerful in that it is easy to use and provides measures that would otherwise require a large amount of time to calculate. We use this tool to determine measures of readability, frequency, imageability, and concreteness, as well as other measures related to semantics. This tool was also used to confirm measures that we had coded manually.

3.2.2 Morphological Analysis

In linguistics and psycholinguistics, a *morpheme* is defined as the smallest unit of language that carries meaning (Whitney, 1998). This is in contrast to a *phoneme*, which is the smallest unit of language, but does not carry meaning and is therefore not of primary focus in this study. In some sense, our morphological knowledge might be thought of as all the words we know as well as the modified versions of those words. However, try to fill in the blanks of these example sentences. If I were to say to you "This is a *wug*. Here is another one, now there are two _____," having a static word list would not allow you to complete the sentence because *wug* is a made up word. Yet, even young children are able to determine that *wugs* would go in the blank. Morphologically, the bound morpheme *-s* is added to pluralize the noun. Now consider another example: "I *arg* my paper. My friend also _____ her paper. We are _____ our papers. Our colleague _____ her paper yesterday." Again, a static wordlist would not allow you to fill in these blanks because *arg* is a made up word. Yet, you were most likely able to determine that *args*, *arging* and *arged* are the correct versions of this novel word. Perhaps even more interesting is our ability to then say "I am an *arger* by profession" in which the verb *arg* has been transformed into a noun.

Thus, it is not a wordlist but rather a set of rules that allows us to combine various parts of words and create new words. Recall that productivity is an essential part of language and it is these morphological rules that in part contribute to productivity. Morphological rules, however, are language-specific. For example, verbs in English have four forms, as can be seen with the example *arg*, while Spanish verbs have about 50 forms (Pinker, 1994). Thus, we have included morphological properties in the analyses because we think it is reasonable to expect that language background and morphological features might interact in a way that has an effect on responses.

Below we describe the sequential process used for assigning morphological properties and features. Table 3.1 provides an example of how we went about coding the exemplar item

1. Part of Speech (POS): As a starting point, each word in each item was categorized into a part of speech category: noun, pronoun, verb, adjective, adverb, determiner, or conjunction. This was verified using the Xerox morphological analysis tool. (Note: parts of speech could fall under the domain of syntax or semantics as the part of speech has to do with the function they have in a structural sense as well as the potential meaning that can be extracted. However, we have included POS as a morphological property since certain morphological rules are determined by the part of speech.)

2. Lemma: In order to use some of the other linguistics tools available for analysis, we had identify the *lemma*, which is the dictionary or canonical form of a word. For example, the lemma for *felt* is *feel* because it is the uninflected form. The lemmas

were first determined by hand and then confirmed using the Xerox Morphological analysis tool.

3. Content or Function: Each word was categorized as to whether it was a content word or a function word. Generally, function words have a syntactic or structural function, like prepositions, auxiliary verbs, determiners, conjunctions, or pronouns. A function word has little lexical meaning or the meaning is ambiguous. Function words primarily serve to express grammatical or structural relationships in the sentence. Within a particular language, function words typically consist of the categories of words that fall under the closed-class grouping, which is a grouping to which new words are rarely added. In contrast, function words consist of categories that are in the open-class grouping (i.e., nouns, verbs, adjectives, and many adverbs). Open class categories accept new words. Words were coded based on their POS.

4. Inflection and Derivation: Inflection is the process of adding morphemes to a lemma so that the role of the word is changed, but the essential meaning is the same. For example, the addition of *-ed* to the end of a verb to make it past tense is a type of inflection. The process of derivation is when the word class and/or meaning of the word is changed as the result of a morphological change. In our exemplar, *hopeful* is a derivation of the noun *hope* plus the morpheme *-ful*. The addition of the morpheme changes the meaning of the word and also changes the part of speech from a noun to an adjective. We coded each word in the item for inflection as well as derivation.

5. Bound and Free Morphemes: As the name suggests, a *free morpheme* is one that can stand alone as a word. These can simply be though of as the lemmas. In contrast, a *bound morpheme* is one that cannot stand on its own, such as the plural *-s* that attaches to nouns.

   Each word was dichotomously coded as to whether it had a bound morpheme or not. For our exemplar, *hopeful* has the bound morpheme *-ful*. Because derivational rules allow us to combine multiple free morphemes into compounds, it is possible for a noun to have two free morphemes. For example, the word *lighthouse* consists of two free morphemes that have been put together.

6. Affixes: Each word was coded for affixes, which are derivational bound morphemes that can occur at the beginning, middle, or end of a word. We coded for prefixes and suffixes. For the exemplar, *hopeful* has the adjective suffix *-ful* added to the noun *hope*, which then changes the category to an adjective. Affixes are bound morphemes by definition, however, given that both suffixes and affixes appear in this questionnaire, we separated out prefixes and suffixes.

These analyses were then summarized and counted to determine the item features. For example, for *I felt hopeful about the future*, there were 6 total words/free morphemes, 5

unique lexical categories, 3 content words, 1 bound morpheme, 1 derived adjective featuring an adjective suffix, and 1 inflected verb.

Table 3.1: Exemplar Reference Table for Coding Morphological features

	I	**felt**	**hopeful**	**about**	**the**	**future**
Lexical Category	N	V	ADJ	PREP	DET	N
Characters	1	4	7	5	3	6
Count of Unique Lexical Categories	1	1	1	1	1	0
Content or Function?	0	1	1	0	0	1
Lemma	I	feel	hopeful	about	the	future
Bound Morpheme	0	1	1 (suffix)	0	0	0
Free Morpheme (word count)	1	1	1	1	1	1
Derivation	0	0	1	0	0	0
Inflection	0	1	0	0	0	0

3.2.3 Syntactic/Structural Analysis

To move from the word-level to the item level, we combine words into sentences. Clearly, words cannot be combined in any order, they are combined according to certain rules. These rules account for the *syntax* of a language. Syntax is able to account for the greatest amount of productivity in a language as forming novel sentences occurs with high frequency. Note that syntax is different from prescriptive grammar rules which are taught in English classes. A run on sentence would be permissible based on syntactic, or grammatical rules, but it would not be permissible by prescriptive grammar rules. We used X-Bar theory, developed by Noam Chomsky, to build our syntax tree structures for each item. This is not the only approach to syntax, but the Chomskian theories represent an approach to identify rules which occur across languages, or a *Universal Grammar*. A syntax tree accounts for the structure of utterances and is guided by particular rules about how words can combine. Students and faculty from linguistics were consulted when drawing the syntax trees to confirm that the correct structure was being represented. Here we briefly describe the steps that were used to code features that emerged from the syntactic diagrams.

1. Syntax Trees: Each item was drawn as a syntax tree to further analyze differences between the items. X-bar schema was used to draw these trees. These trees were compared to the output provided by the Enju parser. In some cases, there was disagreement between our trees and Enju's trees. We provide a syntax tree for our exemplar below, see Appendix for more information about interpreting syntax tree diagrams.

2. Auxiliary and Modal Verbs: A main verb acts as the head of a verb phrase and expresses the main action or state of being. In other words, main verbs contain content. Another type of main verb is the copular verb, which is sometimes called a linking verb or state-of-being verb. In our exemplar, *felt* is an example of a copular verb. The English verb *be* is the most common copular verb. Main verbs and copular verbs are labeled as content words.

   In contrast, an auxiliary verb is an additional verb that adds functional or grammatical meaning. A modal verb is a specific type of auxiliary verb which expresses modality (i.e., likelihood, ability, permission, and obligation) and notably cannot be inflected. about the is Items which contained auxiliary verbs, as identified both in the Xerox tool as well as the Enju tool, were coded and counted for instances where multiple auxiliary verbs occurred. Auxiliary verbs are mainly functional, and are thus labeled as function words. Our exemplar did not contain any auxiliary verbs.

3. Tense and Aspect: The tense and aspect of the verbs was noted and coded based on the syntax trees. Tense is indicated as +/-Past (i.e., Past or not past) and aspect was indicated as either perfect or progressive (i.e., +Perf, +Prog) within a syntax tree diagram. For the exemplar *I felt hopeful about the future*, the verb *felt* is the simple past tense of *feel*.

4. Voice: Items were coded as to whether they contained an active voice construction, a passive voice construction, or both. Our exemplar was worded in active voice.

5. Movement: An aspect of X-Bar theory and other transformational theories of syntax is that there is a deep structure and a surface structure of any sentence. This idea came about as a way of dealing with certain discontinuities, such as passive voice. The basic idea is that certain sentences have the same deep structure, but their surface structures may differ as the result of movement. For example: "She used the computer" is in active voice, the subject (she) occurs before the verb and the object (the computer) occurs after the verb. In contrast, "The computer was used _____ (by her)" has the object (the computer) in the subject position. A transformational syntax theorist would say that these two sentences have the same underlying structure, but that movement has occurred in order for the passively voiced sentence to occur in the surface structure. Thus, we coded for items that showed any movement between the deep structure and the surface structure. Our exemplar did not include any movement.

6. Ambiguity: The technical definition for syntactic ambiguity is a sentence that could possibly have two syntactic structures, leading to different semantic interpretations. Here, we have extended the definition of syntactic ambiguity to include cases where the underlying structure of an item was unclear, even if the meaning was not necessarily different. For example, in the item *I did not feel like eating*, the structure could be *I*

did not feel like eating [food], implying that a nominal category has been left empty, but is mentally accounted for, or if the structure is treating *eating* as a noun, i.e., "the act of consuming food". Therefore, these types of ambiguities were coded. Our exemplar did not include any ambiguities of this type

7. Negative: If there was a negation in the item, this was noted. Our exemplar did not include any negative adverbs.

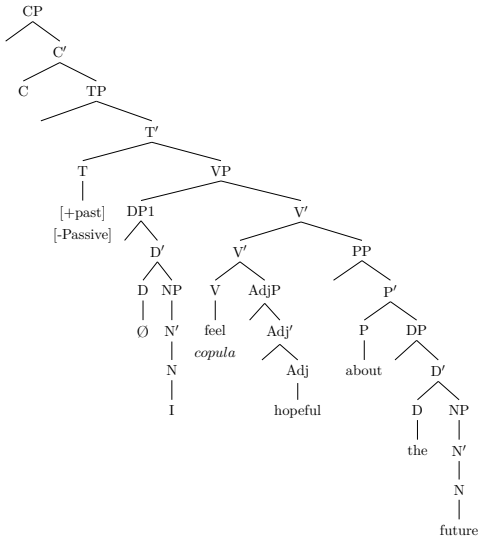
8. Idiom or Collocation: Collocations and Idioms are related concepts, but have proved difficult for linguists and lexicographers to define unambiguously. An idiom is a rigid word combination that cannot be easily broken up and that has a meaning that might be difficult to derive from its respective parts (e.g., Mckeown & Radev, 1999). A collocation is two or more words that occur together more often than would be expected by chance, similar to an idiom. The difference is that collocations can be somewhat understood by their constituent parts, but do not follow typical syntactic or semantic rules. One way to identify such constructions is by replacing one of the words in the suspected collocation to determine if it is permissible and has the same meaning. For example, *feel like* is likely a collocation because if you were to replace either word, it would not be permissible.

   For example, for *I did not feel like eating*, replacing *feel* with *desire* or *experience* creates impermissible sentences like **I did not desire like eating* or **I did not experience like eating*. Another example is *I could not get going*, which if you are to replace *get* with a synonym, *I could not begin going* or *I could not commence going*, it sounds awkward to a native speaker, but would have the same approximate meaning. Alternatively, *I could not get moving* and *I could not get running* are syntactically permissible, but their meaning is not exactly the same. *Get moving* might be interpreted literally as having trouble physically moving about and *get running* might be interpreted the same as *get going* in the context of a machine or computer, but it would be somewhat peculiar to include on a questionnaire. On the other hand, **I could not get starting* is not permissible.

   We considered idioms to be any word segments that could not be broken apart and had a meaning that could not be extracted from their constituent parts. Our exemplar item did not include any collocations or idioms.

9. Syntactic Complexity: Coh-Metrix provides several measures for syntactic complexity. For most of these measures, however, our items had no variance. One of the measures that did have variance across items was the mean number of modifiers per noun phrase. This measure was included in the analysis.

Figure 3.1: Syntax Tree for Exemplar



3.2.4 Semantic Analysis

For aspects related to the meaning of the item or its constituent parts, we used various online tools as well as manually completed ratings. Features in this section relate to the processing of meaning as well as the processing of meaning within a context, which is the domain of pragmatics. Table 3.2 shows the results of the semantic analysis for the exemplar.

1. Subscale: We used the CES-D subscale classifications which were somatic symptoms, depressed affect, positive affect, and interpersonal problems. Our exemplar was on the positive affect subscale, and was thus a reverse-coded item.

2. Behavior or Feeling: Because the subscales which have been identified do not necessarily differentiate between behaviors and feelings, we also coded for this. For example, "I felt that everything I did was an effort" is coded as a somatic item, but we interpreted this item as describing a feeling.

3. Polysemy: As has been mentioned, an ambiguous word is thought to be a word which has many senses, or a *polysemous* word. We used both WordNet as well as FrameNet to count the number of senses for each word and then take an average for the item. We looked up each word in the item, even if it was a function word. To be included as a potential sense, the word had to have the same POS as it did in the item. For some cases, words like *don't* were separated into two parts, a verb *do* and an adverb *not*. This was taken into account when doing the various parts of the analysis.

   According to Fontenelle (2012, p. 442), FrameNet and WordNet are "complementary, given that WordNet is less concerned with the description of combinatory requirements

of a lexical item." As such, we included both because we thought that they were assessing different types of ambiguity. FrameNet has less emphasis on enumerating every possible sense of a particular word, as can be seen by the lower counts of senses provided by FrameNet. WordNet, on the other hand, is more like a dictionary in that it offers a full range of meanings that a word can take.

Coh-Metrix also uses WordNet in the same way that we have proposed, however, it is not clear from the handbook whether Coh-Metrix is accessing the most up-to-date version of the WordNet database. Coh-Metrix also does not consider words that are categorized as function words. In our research, we have found that even some function words have entries in WordNet.

4. Word Information: Coh-Metrix uses the data from the MRC Psycholinguistics Database to calculate a number of measures that we have categorized as falling under semantics or pragmatics (Coltheart, 1981). We chose to include *imageability*, *concreteness*, and *meaningfulness* measures in our analysis. These measures are all based on participant ratings provided in the MRC database by Coltheart (1981), then averaged for the item. Concreteness is defined as how concrete or non-abstract a word is, imageability is the ease at which the mental image of a word can be constructed in the mind, and meaningfulness is how meaningful the word was rated as being, an admittedly unclear definition. Each of these measures ranged from 100-700, with higher values reflecting higher levels of that concept.

5. Hypernyms: Hypernyms are thought to be a measure of abstractness. Similar to the theory of Cognitive Grammar which explains meaning as existing at different levels of abstraction, a hypernym is a superordinate concept which subsumes the word in question. In other words, a hypernym is a more general level of meaning. The hypernym count provided by Coh-Metrix again uses WordNet and examines the conceptual taxonomic hierarchy that is above the given word. The example provided by Graesser et al. (2004) is as follows "For example, *chair* (in the sense of seat) has seven hypernym levels: *seat–>furniture–>furnishings–>instrumentality–>artifact–>object–>entity.*"

6. Frequency: Coh-Metrix provides corpus-based measures for word frequency, or how often words appear in spoken or written language. Because more frequently occurring words are processed more quickly, it is reasonable to include this measure in the analysis. Coh-Metrix uses the CELEX corpus to calculate the mean frequency. Coh-Metrix calculates the frequency by using the frequency count for each word in the item, taking an arithmetic mean, and then taking a logarithm. Coh-Metrix also provides separate calculations for content words and function words and Graesser et al. (2004) consider the mean logarithm of word frequencies for content words to be the primary measure for word frequency.

7. Manually Rated Items: In addition, using four trained raters who were either graduate students or faculty, we coded the items in terms of whether they were *concrete/abstract* as well as *specific/general*. These ratings were done independently and then discussed to come to a consensus coding. In addition, we included ratings about whether the item was *culturally loaded*, or whether we might expect responses to differ in terms of *gender*. We included these ratings in order to see if they were correlated with any other measures and whether there were any differences seen with regards to gender or English fluency.

Table 3.2: Semantic Analysis for Exemplar

Feature	**Value**
Subscale	Positive Affect
Coh-Metrix Concreteness	324
Coh-Metrix Imageability	363
Coh-Metrix Meaningfulness	502
Coh-Metrix Polysemy for Content Words (mean)	3.5
Coh-Metrix Hypernymy Nouns (mean)	6.333
Coh-Metrix Hypernymy for Verbs (mean)	1.462
Coh-Metrix Hypernymy Nouns (mean)	1.299
Coh-Metrix Frequency, content words (log mean)	2.003
Coh-Metrix Frequency, all words (log mean)	2.822
Manual: Behavior or Feeling	Feeling
Manual: FrameNet Lexical Units	2.75
Manual: WordNet Senses	5
Manual: Concreteness	Abstract
Manual: Specificity	General
Manual: Cultural Loading	Not Culturally Loaded
Manual: Gender	Might differ across gender

3.2.5 Descriptive Features

Finally, some descriptive features were included in the analysis which were essentially features that did not fit neatly into another linguistic category. These features included the

length of the item (in terms of count of words, characters, and syllables), as well as the average length of the words in the item (in terms of characters and syllables).

3.2.6 Correlational Analysis

We expect certain features to be dependent in ways which we may or may not be able to anticipate. Thus, bivariate correlations were obtained for the feature codings across the 20 items to determine where dependencies exist. In addition, we are especially interested in seeing how certain item features are correlated with the semantic features.

We first examined the correlations among all the features. This will allow us to easily rule out any features that share considerable overlap. We expect many of the features to be confounded, but it is possible to rule out the most confounded features from the beginning. This helped guide decisions about which features to include in models in order to avoid redundancy. Assigning descriptive, accurate names to our features and properties is one way to make them more user-friendly. Moreover, examining the interrelationships between our item features could provide interesting conclusions about how item properties are related.

3.3 Item Response Models: Probing Person Effects, Item Features, and their Interactions

In this section, we describe our analytic approach for addressing the following research questions:

> **2.6.2 Are linguistic properties/features related to higher/lower endorsement of depressive symptomatology? Which ones?**
>
> **2.6.3 Which properties are most related to higher/lower endorsement of depressive symptomatology?**
>
> **2.6.4 Does language background and/or sex interact with certain item properties to lead to higher/lower endorsement of depressive symptomatology?**

Participants This study was conducted according to Simon Fraser University's ethics guidelines and received Human Subjects Approval from the institutional review board. Participants included undergraduate students recruited through the Department of Psychology Research Participation System (RPS). Participants received credit toward their undergraduate psychology course for their participation. There were no exclusion criteria and the only inclusion criterion was willingness to participate. The sessions lasted approximately one hour, during which the participants responded to a battery of questionnaires related

to health and well-being. Questionnaires were completed both on computer and pencil and paper. This project will be conducting a secondary analysis of the data.

Measures In addition to the CES-D, a demographics questionnaire was administered to participants. Because the individual characteristics we considered were demographic variables, we discuss them in this section.

Demographic information collected from participants included sex, age, ethnic/racial identification, and self-reported level of English Language fluency. Individuals were able to select a single option for their ethnicity, but could provide multiple responses to subsequent questions about their ethnicity. Individuals were able to choose male or female for their sex. For some items, participants were able to choose a non-response option (e.g., sex), and were able to leave any item blank if they chose. Age is an open-ended item which allowed user input. The item measuring level of English fluency asked participants to "Please rate your level of English Fluency" and the response options included "Very fluent, English is my first language," "More fluent in English than in my first language," "Same fluency as my first language," and "Less fluent in English than my first language."

Procedure Participants completed a battery of questionnaires in a lab at Simon Fraser University. Participants completed these questionnaires on either a laptop computer with a mouse, or with pencil and paper. After the participants provided their consent, a research assistant read the instructions for the study. The battery of questionnaires included measures related to health and well-being, and some of the same measures appeared in different formats. For the CES-D, only the first format which participants responded to is included in this study, though future endeavors should certainly look at the multiple time points.

3.3.1 Descriptives

Descriptive statistics were calculated and reported in table format. Frequency counts and percentages were obtained for: sex, Race/ethnicity, and level of English Fluency. Standard descriptive statistics (i.e., mean, variance, standard deviation, median, range, skewness, kurtosis) were computed for age, English fluency, and the composited CES-D scores for the entire sample. In addition, descriptive statistics were provided by sex, and level of English Fluency. Level of English fluency was also presented by Race/Ethnicity. Histograms are provided for visual interpretation of the distribution of CES-D composite scores.

3.3.2 Linear Mixed-Effects Modeling

The analyses used linear mixed-effects models (LMMs) to explore the research questions. Notably, these models will treat the CES-D response as a quasi-continuous outcome variable.

The linear mixed modelling framework takes into account fixed predictors as well as individual-specific random predictors, allowing for item-level analysis. We are treating the

responses as repeated measures, thus, each individual has a vector of 20 responses, each response is on a 4-point scale. A general notation of this model based on De Boeck and Wilson (2004) is

$$Y_{pi} = \sum_{k=0}^{K} \beta_k X_{ik} + \sum_{j=0}^{J} \theta_{pj} Z_{ij} + \varepsilon_{pi} \tag{3.1}$$

where k represents an index for predictors with a fixed effect (in our case, item features, sex, level of English Fluency); j represents an index for predictors with a random effect (e.g., person); X_{ik} is the value of predictor k for item i (for k=0, X_{i0} for all i);

Z_{ij} is the value of predictor j for item i (for j=0, X_{i0} for all i); β_k is the fixed regression weight of predictor k, an overall intercept for k=0 and predictor-specific effects for k=1,....,K; θ_{pj} is the random regression weight of predictor j for for person p, a person specific intercept for j=0, and person-specific slopes for j>0; and ε_{pi} is the error term for person p and item i.

The assumptions for this model include:

1. All relevant fixed predictors are included in model.
2. All relevant random predictors are included in model.
3. Each random effect is normally distributed and jointly follow a Multivariate normal distribution.
4. The error terms, $\varepsilon_{\mathbf{pi}}$ follow a normal distribution with a mean of 0.
5. The error terms for the random effects jointly follow a multivariate normal distribution (per person) such that $\varepsilon_p \sim N(\mathbf{0}, \mathbf{\Omega})$, where $\mathbf{0}$ represents a vector of zeros, $\mathbf{\Omega}$ denotes a covariance matrix of the error terms for each item (20 x 20).
6. The covariance structure of the random effects is properly specified.
7. The fixed predictors do not covary with the errors or the random effects.
8. Sufficient sample size.
9. Missing data are assumed to be missing completely at random (MCAR).
10. While not a formal assumption, $\mathbf{\Omega}$ is often assumed to be a diagonal matrix (i.e., none of the items covary with other items; assumption of local independence), and sometimes, the item variances are assumed to be equal. However, this covariance structure often does not provide the best fitting model. Steps will therefore be taken to ensure the appropriate covariance structure is used.

Violations of (2), (3), and (4) and (5) do not often impact the coefficient estimates, though diagnostics used to test assumptions will be described below.

Model Fit and Covariance Structure for Within-Subject Residuals While there is no universal approach to selecting the proper covariance structure, given that linear mixed models are used in a wide variety of situations, there are some general recommendations. Selection of an inappropriate covariance structure can bias parameter estimates, and in some cases, use more degrees of freedom than necessary.

The literature suggests that for repeated measures data in which the measurements were taken in quick succession (as opposed to longitudinal data, where time points are often months or years apart) would lend itself to some sort of autoregressive model (e.g., Vonesh & Chinchilli, 1997). Autoregressive covariance structures are patterned in such a way that measures which are closer together in time have larger covariation which decays at some predefined rate. For our purposes, this would mean that we would expect item 1 and item 2 to have a larger covariance than item 1 and item 20. Therefore, a first order autoregressive structure, Toeplitz (banded), or ARMA (Autoregressive moving average) might provide the best fit. The heterogeneous variants of these models may also provide acceptable fit indices.

Another possibility is a compound symmetric covariance structure, which constrains the off-diagonal elements to be equal. While it is certainly possible that this structure would fit our data, previous research from our lab has shown that, across the 20 items, the responses tend to decrease. Therefore, observing a negative slope from item to item, rather than a flat line relationship, certainly would point to an autoregressive structure rather than a structure which assumes equal covariances for items, regardless of how far apart they are.

We did not expect that a highly constrained identity covariance structure, which assumes variances equal to one and off-diagonal elements set to zero, to be appropriate. Conversely, we also did not expect that an unstructured covariance structure, which has the fewest constraints and allows for every element in the matrix to be different, to be appropriate. An unstructured covariance specification is also costly in terms of computational resources and degrees of freedom since so many parameters are being estimated (especially for 20 measures per participant).

In order to determine the covariance structures, we tested a model which included English fluency, sex, and multiple item features with each potential covariance structure specification. We used Akaike's Information Criterion as well as Bayes' Information Criterion (which penalizes for model complexity) to select the best fitting structure for each model.

Proposed LMMs

1-Predictor Main Effects Models Each feature was included in a single predictor model. Features that were not statistically significant in single predictor models were excluded from further analysis

3-Predictor Models: Feature with English Fluency and Sex A model which included the linguistic feature, English fluency, and sex, was estimated. A symbolic representation of these models would appear as:

$$\hat{y} = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 W + \varepsilon$$

Where $\hat{y}$ is the outcome variable, the CES-D item response; β_0 is the intercept; β_1 is the fixed regression weight of predictor X, which represents the levels of the item feature; β_2 is the fixed regression weight of predictor Z, which represents the level of English Fluency; β_3 is the fixed regression weight of predictor W, which represents the sex; and ε is the error term. The three partial main effects were evaluated using F-tests, and t-tests of parameter estimates were interpreted.

If all three partial main effects were not statistically significant, no further models were obtained for the feature being examined. The purpose of these models was to determine if the feature appeared to have an effect when controlling for English fluency and sex.

4-Predictor Interaction Models: Feature with English fluency, Sex, and Interaction Two 4-predictor models were estimated for selected features. These models build on the 3-predictor models by including a 2-way interaction term. One model included the feature by English fluency interaction term and the second model included the feature by sex interaction term. The partial main effects and interaction effects were interpreted using F-tests and coefficient estimates were evaluated using t-tests. Models which did not have statistically significant interaction effects were not included in subsequent analyses. These models were used to determine if the effect of the feature varied across sex or levels of English fluency when controlling for all three partial main effects. As is typically the case for models with interaction terms, the interaction was of primary interest in these models.

5-Predictor Model: Feature, English fluency, Sex, and Interaction terms A 5-predictor model was estimated for selected features and included the feature, English fluency, sex, feature by English fluency interaction term, and feature by sex interaction term. This model examined whether the effect of the feature was different depending on sex as well as English fluency.

Summary and Limitations A mixed-modelling approach allowed for examinations of the effects of item features and individual characteristics on responses, as well as the interaction between either type of variable. However, there is a major limitation of this approach. In this framework, the outcome variable (CES-D response) is assumed to have a continuous distribution. Since the CES-D response scale is four points, interpretation of results from a linear mixed model might be problematic. When composite scores from the CES-D are used, there is a stronger argument for using a linear model. For items with four response

options, item-level analysis can potentially cause problems. To begin, a linear model might make predictions outside of the range of possible responses. It would not seem to be the case that the response options are nominal, assuming that there is order in the response options is defensible. However, making the case for an interval or ratio scale is challenging as it is unclear if there is additivity among the response choices.

Moreover, the response options do specify counts of days. Count variables are ordered and additive, but do not typically follow a symmetric distribution. Therefore, there is rationale to move outside of the general linear model framework.

On the other hand, there are advantages of using linear mixed effects models. To start, the interpretation of the coefficients from these models is often more simple, especially when higher order effects such as interactions are included.

3.3.3 Generalized Linear Mixed-Effects Modeling

Because the current project is exploratory, we also formulated selected models within a generalized linear mixed model framework. GLMMs were used because they allow for the outcome variable to have an error distribution that is non-normal. Because the CES-D responses for single items are bounded at 0 and 3, a linear model is likely to predict values outside of this range.

There are a variety of generalized linear mixed models that we could potentially use. In order to select the most relevant, we must consider three specifications which will build the models we test:

1. The distribution of the data: We conducted exploratory data analysis and non-parametric tests to determine which distribution our response data are likely to be drawn from. We expected to model the CES-D responses using a Poisson and/or Multinomial distributions.

2. The link function: The canonical link functions were used for each type of distribution modelled. For a Poisson distribution, we used a log link function.

3. The predictors: We used the same predictors described in the previous section – English Fluency, Sex, Features, and interaction terms (as necessary).

Features that we selected for inclusion in the GLMMs were based on the results from the linear mixed model analyses. We chose features that appeared to have large effects on the outcome variable of the CES-D response, features that had moderate effects, and features that had small effects. The aim was to include a feature from each domain and ensure both positive and negative effects were included. If there is sufficient evidence that these models offer a better fit to our data, as assessed by the AIC and BIC, additional models were obtained for other features.

3.3.4 Diagnostics and Type I and Type II Error Control

Assumption Checking We tested the assumptions for our models using standard diagnostic procedures. To ensure a linear relationship, a plot of the residuals with the predicted CES-D score was fitted with a Loess line and visually analyzed. A normal q-q plot of residuals was visually assessed to check for non-normality, as well as the Kolmogorov-Smirnov test for normality. Multicollinearity among predictors was assessed using the variance inflation factor (VIF).

We did not expect to find any extreme values since the response variable is restricted. In terms of our predictors, we did not expect there to be extreme values, nor did we want to exclude any items if they were unlike the other items. We did not exclude data from any participants based on any indices such as missing CES-D data, extremely high or extremely low composite CES-D scores, or suspicious response patterns.

Type I and Type II Error Control As an exploratory study that tested many different models, the control of type-I errors was of particular concern. With so many tests, it was highly improbable that we would *not* find at least one test to be significant. A first step toward type-I error control is examining the relationships among the linguistic features, as this allowed us to reduce the number of predictors, allowing for fewer tests. In addition, we reported all the tests that were conducted (other than those used to select covariance structure), rather than a select subset of the tests conducted. While this does not control for type I error, this transparency allows for the results to be interpreted more accurately.

A highly conservative approach we considered was to keep a running count of the number of models tested (in other words, treating the entire study as a "family" of tests) and then use that number to perform a Bonferroni correction to keep the family-wise type I error rate to α=.05. However, there are several reasons why this study-wide level of control is not desirable. To start, it does not take into account dependencies between various tests, and also, that the interpretation of a given test can change simply by increasing the number of tests. Intuitively, this does not make much sense; statistically, it reduces the power of the study. In sum, a study-wide Bonferroni correction to control for type-I error is too conservative and inappropriate.

Thus, we chose an approach which allowed us to control for type-I errors, without being overly conservative, but still controlling at the family-level rather than the individual test-level. In order to do this, we categorized the proposed models into families of tests that have similar hypotheses.

To assess power, we used the program G*Power 3.0.10 to calculate the sample size needed to conduct F-tests using a repeated measures design that included two predictors as well as an interaction term. We specified an α=.05 and power (i.e., the probability of detecting a deviation equal to or larger than specified effect size) to be .95, 4 groups, and 20 repeated measures. To detect a small effect size (i.e. f(V)=.1), we would need a sample

size of 1524. However, it is worth noting that as a secondary analysis, the utility of a power analysis is limited.

3.3.5 Software

SPSS version 19 and version 24 were used for analyses. Example syntax for these analyses is included in the appendix. Estimates were obtained using a maximum likelihood estimation algorithm.

3.3.6 Interpretation

The use of standardized effect sizes can sometimes obscure the meaning of results. Thus, we interpreted our results in terms of the original metric for each feature. We considered that the CES-D is typically composited when interpreting the results. The regression coefficients were interpreted for all of the partial main-effects models when they were statistically significant. Interactions were examined on a case-by-case basis.

As is common in exploratory studies, we did not establish any specific expectations for a particular type or magnitude of standardized effect size. We were generally interested in effects of any size. However, given that the CES-D is used frequently and its utility as a depression measure has been established, we did not expect that there would be large effects for linguistic features. That is, we did not think that the CES-D responses would be as strongly related to linguistic features as they were to depressive sympotmology. However, as a guideline, we consider any feature which results in a change of one point on the composite score to be a moderate effect, and a feature which results in a change greater than one point to be large.

Broadly, positive values for coefficients in linear mixed models represent an increase of depressive symptomology when holding other predictors constant and negative values represent a decrease. Because of the way English Fluency was coded, a one unit increase in the variable corresponds to a *decrease* in English fluency. That is, the scale ranges from 1=Very fluent in English, English is my first language to 4=Less fluent in English than in my first language. Therefore, we interpreted a one unit increase for this variable as an increase in non-English dominance, or English as an additional language. We preferred this framing as opposed to other possibilities such as "English novelty," which suggests some temporal relation to English, or "lack of English fluency" which seems to have a negative connotation.

Interpretation for generalized linear mixed models is inherently more complex. For models where the outcome is treated as a count and there is a log link function, the coefficients are given in terms of log counts. Log counts can be additive, but are not as meaningful for interpretation because a log count is no longer in the metric of the outcome variable. The log count can be exponentiated to provide a multiplier for the expected counts. How-

ever, this nonlinear transformation means that the effects are no longer additive and can be multiplicative. Thus, interpretation for a exponentiated coefficient depends on all other predictors being held constant at particular values. Thus, the meaning of a coefficient for a continuous predictor if we assume a log link is as follows: for a one unit increase in the predictor, the log CES-D response increases/decreases by the value of the coefficient. For the expected counts, the interpretation is as follows: each additional one unit increase in the predictor multiplies the CES-D response by the exponentiated coefficient. Because the exponentiated counts will always be positive, we interpret values less than 1 as corresponding to decreases in CES-D responses and positive values as corresponding to increases in CES-D responses.

Chapter 4

Results

4.1 Correlational Analysis

4.1.1 Procedure

To address research question 1, we used descriptive and correlational analyses to examine variables. The purpose of this approach was two-fold: The first reason is mechanical; we were starting with a large set of variables which we expected to carry numerous dependencies and it was, thus, important to select a subset of these variables which had enough variability across the items to be useful, and that might contribute uniquely to our models. The second reason is substantive, that is, we would be able to examine how the variables were related across linguistic domains. In particular, we were interested in how the semantic features might be related to features in other domains.

The first step in our analysis was to gather descriptive statistics (i.e., minimum, maximum, mean, standard error, standard deviation, and variance) for the subset of features within each linguistic domain. We were looking for features that had the largest amount of variability. Then, we obtained bivariate Pearson Product Moment Correlation (PPMC) coefficients and conducted two-tail significance tests for each feature. Based on these results, a preliminary set of features within each domain was considered.

4.1.2 Length Features

Our descriptive properties included various ways of measuring the length of an item. The specific features were:

1. The count of words in the item
2. The mean number of syllables for each item
3. The standard deviation of syllables for each item
4. The mean number of letters per word for each item

5. The standard deviation of letters per word for each item

Descriptive statistics appear in table C.1. The total number of words in the item had the highest mean (($\bar{x} = 6.3$) and the most variability ($sd = 3.77$ words). The average number of syllables per item had a mean of 1.30 and a standard deviation of .2. Syllables generally have some variability in how many letters are present, which means that an item that has a variety of long words and short words will not necessarily have variability in the number of syllables. This was evident by the maximum for average number of syllables per word per item being 2 and the small standard deviation. The average number of letters per word had a standard deviation of .80, with the average number of letters per word being approximately 4 letters. This means that words tended to be short. For the feature of standard deviation of the number of letters per word for each item, the mean was 2.36, though we see that the maximum was 4.04. This means that for at least one item, there was considerable variability in the number of letters per word. Based on these descriptive statistics alone, number of words and average number of letters per word were considered to be optimal features, given their variability. We did not think that the two measures for syllables would be as informative to include in the models because they vary in number of letters

Correlations appear in table C.2. These results indicate that the word count feature does not strongly correlate with any of the other length features except the average letters per word per item ($r = -.45, p < .05$), confirming our selection of this variable. Additionally, the two measures for syllables are strongly intercorrelated, as well as strongly correlated with the two features for number of letters per word per item, again, supporting our choice to exclude these variables. Next, the average number of letters per word for each item is strongly correlated with the standard deviation for letters per word ($r = .59, p < .01$). Thus, we select the number of words per item and the average number of letters per word per item for inclusion in further analyses. The two features regarding syllables are eliminated due to (1) lack of variance and (2) high intercorrelations with other variables which are more reasonable to include.

4.1.3 Morphological Properties

Next, we considered the morphological features that we coded. These included:

- Counts of the parts of speech (POS) categories (i.e., noun, main verbs, copular verbs, auxiliary verbs, modal verbs, pronouns, adjectives, adverbs, complements, prepositions, and determiners).
- Proportion of [POS] to the total number of words per item.
- Number of content and function words; proportion of content or function words to total number of words per item.

- Counts of inflections and derivations; proportion of inflections/derivations to total number of words per item.
- Counts of bound and free morphemes; proportion of bound morphemes to total number of words per item.
- Counts of prefixes and suffixes; proportion of prefixes/suffixes to total number of words in item.

Morphological Properties I – Parts of Speech

Because the part of speech features are somewhat separate from the other types of morphological features, we have separated our morphological analyses into two parts and this will be the case when we are testing models as well.

Results of our morphological analyses begin in table C.3, which includes the descriptive statistics for the POS variables coded as counts. The verb category (i.e., total count of main verbs, copular verbs, auxiliary verbs, and modal verbs) had the largest mean and standard deviation ($\bar{x} = 1.9; sd = 1.12$), suggesting that verbs are the most common POS and the most likely to vary. Further inspection of table C.4, which has the relative frequency/proportion, showed that there is less variability, compared to the other features, when considering the count of verbs relative to the total count of words ($sd = .10$). Note that the verb category is the only category which appeared at least once in every item (as indicated by the minimum). That seems reasonable since a verb is necessary to form a sentence. The variability for main verbs (V_Main) is larger relative to the other categories when coded as a proportion ($sd = .13$). This suggests that the number of main verbs does not neccessarily vary proportionately with the length of the item. A similar pattern can be observed for adjectives, which show more variability relative to the other features when coded as a proportion ($sd = .15$) than as a count ($sd = .59$). Pronouns, on the other hand, have less variability relative to the other features when they are coded as a proportion ($sd = .09$) than when they are coded as a count ($sd = .99$). This suggests that the proportion of pronouns stays relatively consistent regardless of how long the item is.

Nouns tended to vary approximately the same amount relative to the other features regardless of whether they were coded as counts ($sd = .85$) or proportions ($sd = .12$). This is also the case for other categories such as prepositions, determiners, adverbs, complementizers, and the other verb categories. Thus, preliminarily, we would select nouns, main verbs, pronouns, and adjectives for our models. This is an interesting result as these categories, except for pronouns, are typically considered to hold content. Therefore, we can tentatively conclude that content is more likely to vary from item to item than function words.

Next, we obtained the correlations among the parts of speech features for both counts and proportions. We will first discuss the counts. Notably, pronouns and total verb counts are correlated at .86 ($p < .01$), and similarly, main verbs and pronouns are correlated at

.80 ($p < .01$). Because pronouns were overall more prevalent than nouns, and an item must have either a pronoun or a noun, plus a verb, these strong correlations are not surprising. Similarly, nouns are correlated with verbs, pronouns, and prepositions. The correlations range from .6-.64 ($p < .01$), which suggests that nouns, pronouns, and prepositions frequently co-occur, though pronouns and prepositions co-occur less frequently than nouns and prepositions. Adverbs and prepositions were correlated at .73 ($p < .01$), though it is not immediately clear why this is the case. Additionally, verbs and adjectives ($r = -.62; p < .01$) as well as main verbs and adjectives ($r = -.65; p < .01$) are negatively correlated, but adjectives and copular verbs are positively correlated ($r = .27; p > .05$), though this was not a stastistically significant correlation. This is reasonable considering that adjectives did not occur in every item and when they did occur, it was in the form of "I was [adjective]."

Compared to the table showing correlations among proportions of the various POS, the correlations are generally lower and/or non-significant. This is with the exception of main verbs and copular verbs, which were more strongly correlated ($r = -.78$; $p < .01$), and total verbs and modular verbs ($r = .59$; $p < .01$).

The meaning of the correlations for the proportions is whether the *weight* of the occurrence of one POS corresponds similarly to the weight of another POS. We expect that that the longer the item, the smaller the relative frequency or weight will be for any given POS. For instance, when examining the correlation between the relative frequency of nouns and the relative frequency of verbs, we are asking whether higher relative frequencies of nouns correspond to higher relative frequencies of verbs. Based on our results, we decided to use the proportion features rather than the count features. The rationale is two-fold: we do not want to have redundant or highly correlated features in our analyses, and we want our results to potentially generalize, which is complicated when using counts since the count of a particular POS is contingent on the total length of the item.

Morphological Features II

The descriptive statistics for counts (frequencies) of various types of morphemes and morphological processes are found in table C.7, and the descriptive statistics for the ratios, or relative frequencies, are found in table C.8. In table C.7, we see that the number of function words varies more than any of the other morphological variables ($sd = 2.38$). On average, however, there are fewer function words in an item ($\bar{x} = 2.8$) than there are content words ($\bar{x} = 3.5$). There are relatively few derivations per item, the maximum was 1. There was an average of approximately 1 bound morpheme per item. Inflections were more common, each item had at least 1 inflected word, 4 was the highest number seen. Suffixes ($\bar{x} = .8; sd = .77$) were more common than prefixes ($\bar{x} = .15; sd = .37$) and had more variability.

In comparison, table C.8 reflects similar information in terms of proportions. On average, 61% of the words in the item were coded as content words and, thus, the remaining

39% were function words. There was an average of .21 bound morphemes per item and .28 inflections. Bound morphemes had the most variability ($sd = .17$).

Table C.9 includes the correlations among the counts of the various morphological variables included in this set. The correlation between function words and inflections is quite high, $r = .84$ ($p < .01$), as is the correlation between content and function words ($r = .79; p < .01$) and content words and inflections ($r = .68; p < .01$). Next, bound morphemes and suffixes were correlated at $r = .64$ ($p < .01$), while derivations and suffixes as well as function words and suffixes had fairly large correlations ($r = .50; p < .05$ and $r = .47; p < .05$, respectively).

Correlations for the proportions of these features are provided in table C.10. There are some changes in the magnitude and direction of correlations. For example, the high correlation between function words and inflected words is now ($r = -.19; p > .05$), having switched directions as well. The correlation between function words and suffixes has also changed directions from $r = .47$ ($p < .05$) to $r = -.34$ ($p > .05$). On the other hand, proportions of derivations and suffixes are now highly correlated. The correlation between derivation and content words was $r = -.23$ ($p > .05$) when coded as counts, and is $r = .49$ ($p < .05$) when coded as relative frequencies.

Because the results thus far did not clearly suggest the use of counts or proportions, we diverged slightly from our analysis plan and examined correlations between the length of the item and the parts of speech features. This additional analysis was intended to help answer the following questions:

- Do the relative frequencies of each type POS, or morpheme, stay constant as the length of the item increases?
- The implicit assumption thus far has been that there is a positive linear relationship between the frequencies for POS and word length, is this the case?

Morphological Properties and Length Using the number of words to quantify the length of the item, we obtained correlations between word length and each morphological category. We expected that the total counts of each type of POS would generally correlate positively and significantly with the total word count, and that the proportion for each POS would not correlate statistically significantly. In table C.11, when the POS variables are in terms of counts, there are many statistically significant correlations between each POS and the total word count. Pronouns and prepositions had the highest correlations ($r = .83; p < .01$ for both), followed by the total noun count ($r = .80; p < .01$), total verb count ($r = .79; p < .01$), adverbs ($r = .72; p < .01$) and main verbs ($r = .60; p < .01$).

When considering the POS variables as proportions, prepositions were still highly correlated, though the correlation coefficient was smaller ($r = .62; p < .01$). Additionally, adverbs were still correlated fairly strongly, but not as strongly as when adverbs were coded

as counts ($r = .48; p < .05$). This indicates that prepositions and adverbs were relatively more frequent the longer the item. Neither proportion of nouns ($r = .17; p > .05$) nor pronouns were statistically significantly correlated with the total word count and in fact, pronouns were correlated in a different direction ($r = -.14; p > .05$). Similarly, main verbs and total verb count were not statistically significantly correlated and had negative correlations ($r = -.24; p > .05$ and $r = -.23; p > .05$). Adjectives were strongly negatively correlated with the total word length ($r = -.65; p < .01$) indicating that the proportion of adjectives was lower in longer items.

The results of the correlational analysis are reported in table C.12. Not surprisingly, the total word count was correlated most strongly with the count of function words ($r = .96; p < .01$), followed by the number of content words ($r = .92; p < .01$). Inflections also had a strong correlation ($r = .82; p < .01$). The other morphological features did not have statistically significant correlations.

For proportions, function words continued to have a strong positive correlation with the total count ($r = .68; p < .01$), indicating that the longer the item was, the larger the proportion of function words tended to be. This of course would mean that the relative proportion of content words decreased as the length of the item increased. Similarly to the pattern seen with adjectives, derivations were more strongly correlated when they were coded as proportions than when they were coded as counts. This seems to be related to the types of items that included adjectives; they tended to be very short. The derivations were almost always adjectives.

From these analyses, we chose to keep the following variables: nouns, main verbs, pronouns, prepositions, adverbs, adjectives, auxiliary verbs, and copular verbs. All of the POS features were coded as proportions in the analyses rather than counts. In addition, content and function words, bound morphemes, and suffixes were included in the analyses. Bound morphemes and suffixes were coded as counts.

4.1.4 Syntactic Properties

Descriptive statistics for the syntactic features are in table C.13. These results identify two variables which have no variability (i.e., past tense and active voice). These features were thus removed from further analysis. Ambiguity ($sd = .47$) and negations (.47) had the most variability, followed by collocations ($sd = .44$) and whether the item contained a possessive (.44). Passive voice and perfect tense aspect had the lowest amount of variability ($sd = .22$) as these features only occurred in a single item each. In addition, the progressive tense aspect and present tense only occurred in 2 or 1 items, respectively. Therefore, we will not consider these features in further analyses.

After obtaining the correlations among syntactic features, there were a number of statistically significant correlations. For example, collocations and ambiguity were correlated at $r = .88$ ($p < .01$), and trace and relitivizers were correlated at $r = .79$ ($p < .01$). Col-

locations were also highly correlated with idioms ($r = .58, p < .01$). We chose to include ambiguity instead of collocations because there was less overlap between ambiguity and idiomatic items. Additionally, we chose to remove trace because relativizers/relative clauses are less linked to theory than trace. Additionally, the noun phrase feature was excluded because nouns and pronouns were explored in-depth as a part of the morphology domain.

4.1.5 Semantic Properties

Descriptive statistics for the semantic features are provided in table C.15. Note that while all semantic properties were examined together here, in the linear mixed models, they were separated into manually coded features and features derived from Coh-Metrix. There were three features examined for word frequency, each were a logarithm of the mean of the frequencies based on the CELEX corpus. Higher values for any of the three variables indicate more frequent, or commonly occurring, words. The frequency for all words ranged from 1.62 to 3.53 and the frequency for content words ranged from 1.62-3.52. The results suggest that there is a small difference between the frequency for all words ($\bar{x} = 2.84$) and the frequency for content words ($\bar{x} = 2.64$), while the standard deviations were similar ($sd = .52; sd = .53$). This suggests that most of the words in the items were considered to be content words in this analysis. The minimum frequency, which is had a smaller mean ($\bar{x} = .19; sd = .96$), but more variability. Because this is a logarithmic scale, it is not clear what the units mean absolutely.

For familiarity, concreteness, imageability and meaningfulness, the potential values ranged from 100-700, with higher scores indicating higher levels of that feature. For familiarity, the average was 591.3, with a standard deviation of 14.85. Therefore, the words used in these items tended to be familiar words, and there was little variability. The average for concreteness was 328.5, indicating that the items tended to not contain many highly concrete words, and the standard deviation was 36.03, indicating some variability. Imageability was also relatively low ($\bar{x} = 382.24$), indicating that these items were not very imageable, and the standard deviation was 46.01, suggesting a moderate amount of variability. For meaningfulness, the average was 456.55, indicating that the items contained words that were meaningful quite often, though this feature had the most variability out of the four ($sd = 65.83$).

For the measure of polysemy from Coh-Metrix, which only included content words, the average was 5.24. This means that on average, each content word in each item had about 5 possible meanings, with the maximum being 11.83. For hypernymy, larger values indicate more specificity as they represent superordinate levels of meaning. For hypernymy of nouns, the average was 4.03, indicating that nouns in these items tended to be fairly specific. Verbs, on the other hand, were more general, as they only had 1.45 superordinate levels of meaning on average. The standard deviation for noun hypernymy was 3.15, indicating that there was a large amount of variability for nouns, while the standard deviation for nouns was quite

low at .74. Interestingly, the hypernymy for nouns and verbs was only 1 on average and had a smaller maximum than either nouns or verbs alone ($sd = .54$). This would indicate that items which had nouns that were more specific had verbs which were highly non-specific, therefore, the average was pulled down.

Next, polysemy was assessed in several ways. Manually coded polysemy based on the FrameNet tool as well as the WordNet tool, as well as measures provided by Coh-Metrix which only examined content words. Results showed that measures of polysemy were smaller when they were from FrameNet ($\bar{x} = 3.15$) and larger when they were obtained from WordNet ($\bar{x} = 8.54$). This is not surprising as WordNet does not take the semantic category and role of the word into account as much as FrameNet. So a word that has many meanings in WordNet might have few frames in FrameNet because the different meaning have similar semantic roles. The manual ratings of polysemy were larger than the polysemy ratings provided by Coh-Metrix ($\bar{x} = 5.24$), indicating that so-called function words could actually contain a great deal of potential senses.

The averages for the manually rated features and the subscales represent the proportion of items that were coded as containing the feature. For specificity, items were generally rated as being more general than they were specific ($\bar{x} = .25$, i.e., 5 items were rated as specific, 15 as general). Items were rated as being concrete more often than they were rated as being abstract ($\bar{x} = .65$). Approximately half of the items were considered to be culturally loaded ($\bar{x} = .45$), and fewer items were loaded with respect to gender ($\bar{x} = .35$) than they were with respect to culture. For the subscales, the somatic subscale and depressed affect subscale accounted for most of the items (14 total), while positive affect accounted for 4 items and interpersonal problems accounted for 2. In terms of whether items are behaviors or feelings, 6 items described behaviors and 14 described feelings.

Correlations among semantic variables were obtained and are reported in table C.16. Beginning with the frequency features, the frequency for content words correlated at .86 ($p < .01$) with the frequency of all of the words, confirming that there is a great deal of overlap between these two variables. It is likely that the words that had frequency ratings were more often content words than function words. Next, the frequency of content words had a strong negative correlation ($r = -.64$; $p < .01$) with meaningfulness indicating that the more meaningful an item, the less frequently the words occurred. The frequency of content words had a strong positive correlation ($r = .58$; $p < .01$) with the polysemy measure provided from Coh-Metrix, indicating that more commonly used words also had more possible senses. The frequency of content words was also strongly correlated with the manually-obtained polysemy feature which used WordNet ($r = .59$; $p < .01$), as would be expected. The frequency of all the words was also correlated in the same direction with polysemy provided by Coh-Metrix ($r = .42$; $p > .05$) and manually-obtained polysemy using WordNet ($r = .51$; $p < .05$), but the correlations were less strong. The minimum frequency feature had statistically significant correlations with the Coh-Metrix measure of polysemy

($r = .48$; $p < .05$) as well as the WordNet manually-derived polysemy feature ($r = .47$; $p < .05$).

Familiarity and the FrameNet manually-obtained polysemy measure were correlated at -.49 ($p < .05$), indicating that the more familiar a word, the fewer possible frames it had. In addition, familiarity and the depressed affect subscale had a -.52 correlation ($p < .05$), indicating that items containing more familiar words tended to not be on the depressed affect subscale, however, they did correlate at .53 ($p < .05$) with the interpersonal problems subscale. Concreteness and imageability had a .73 ($p < .01$) correlation, indicating a strong positive relationship between these features. In addition, concreteness and meaningfulness had a .54 ($p < .05$) correlation. Concreteness and items on the interpersonal problems subscale correlated at .75 ($p < .01$), indicating that these items tended to have more concrete words. Similarly, items that referred to others correlated at .53 ($p < .05$) with concreteness. Imageability was positively correlated at .70 ($p < .01$) with meaningfulness, indicating that there is some overlap in the concepts of imageability and meaningfulness. Additionally, imageability was negatively correlated at -.57 ($p < .01$) with the Coh-Metrix polysemy measure. This indicates that the more imageable the words in an item were, the fewer possible meanings they had. Similarly, imageability and the manually-derived WordNet feature for polysemy were correlated at -.51 ($p < .05$). Meaningfulness and the Coh-Metrix measure for polysemy had a strong negative correlation ($r = -.62$; $p < .01$) indicating that the words in an item were more meaningful when they had fewer possible meanings. Similarly, the WordNet manually-derived feature for polysemy had a correlation of -.55 ($p < .05$) with meaningfulness. This is an interesting finding as it suggests that words that are semantically ambiguous are generally considered to be less meaningful. Finally, items that were coded as behavior items had a -.53 ($p < .05$) correlation with meaningfulness.

As expected, the Coh-Metrix measure of polysemy had a high correlation with the manually-derived WordNet polysemy feature ($r = .92$; $p < .01$). The hypernymy feature for nouns was correlated strongly with the hypernymy for nouns and verbs ($r = .84$; $p < .01$), while the hypernymy feature for verbs and the hypernymy feature for nouns and verbs did not correlate statistically significantly with each other ($r = .30$; $p > .05$), or any of the other features.

Next, we will discuss the manually-coded features. The FrameNet count of senses (polysemy) had a statistically significant positive correlation with the depressed affect subscale ($r = .47$; $p < .05$). The WordNet count did not have any statistically significant correlations that have not previously been mentioned. Specificity had a .54 ($p < .05$) correlation with the somatic subscale and a .88 ($p < .01$) correlation with items coded as being related to behaviors. Concreteness also had a .54 ($p < .05$) correlation with the somatic subscale, but had a weaker correlation with items that were coded as being related to behavior ($r = .48$; $p < .05$), though this correlation was still statistically significant. Neither culturally loaded items nor gender related items were correlated with any of the other semantic items, which

reflects the pragmatic aspect of these items (i.e., their meaning within a certain context is being probed). Items that were reverse-coded/on the positive affect subscale were not statistically significantly correlated with any of the other features. Items on the somatic subscale had a negative correlation ($r = -.54$; $p < .05$) with items on the depressed affect subscale. These items were also strongly correlated with the coding of items as being related to behaviors or feelings ($r = .66$; $p < .01$). This was to be expected since the behavior vs. feeling feature was included because some of the items that were considered to be a part of the somatic subscale seemed to be highly related to specific feelings. Depressed affect did not correlate with any of the remaining features. Interpersonal problems had a strong correlation with the manually coded feature of referring to others ($r = .67$; $p < .01$), as we expected. The remaining features were not strongly correlated.

Based on these results, we will retain the following variables: frequency of all words, familiarity, concreteness, imageability, meaningfulness, hypernymy for nouns and hypernymy for verbs, FrameNet count of senses, WordNet count of senses, specificity, concreteness (manual rating), cultural loading, gender loading, all the CES-D subscales, behavior/feeling, and whether the item refers to others. Again, in the linear mixed model analyses, the semantic domain is separated into semantics I and semantics II. Semantics I contains all of the manually rated features (except for polysemy) and the subscales. Semantics II contains all of the features derived from Coh-Metrix, as well as the manually obtained measures of polysemy.

4.2 Descriptive Statistics

4.2.1 Data Screening and Preparation

The dataset used was first screened for any data that should be removed. Since we are looking at the item level, an individual who was missing responses to some of the CES-D items was not excluded. However, individuals who were missing all of the CES-D items were removed. Our original dataset had 2215 respondents, 14 of which were then removed because they were missing CES-D responses for a total of 2201.

A summary for valid and missing data for age, sex, ethnicity, and English fluency is provided in C.17.

4.2.2 Sample

Because there was a single individual who did not indicate their sex, their responses were not included in the analyses where sex was a predictor. Likewise, any individuals who did not report their English fluency were not included in the analyses.

4.2.3 Descriptive Statistics

Descriptive statistics for age, CES-D composite scores, and English fluency (coded as a continuous variable) were obtained and are reported in table C.18. The ages of participants ranged from 17 to 53 years with a mean of 19.8. As would be expected from a college sample, the distribution was leptokurtic and positively skewed, indicating that most of the participants were younger. For the CES-D composite, on which higher scores indicate higher endorsement of depressive symptomology, the highest reported score was 51, which is 9 points lower the the theoretical maximum of 60. The average was 15.17, though there was quite a bit of variability. The distribution had a positive skew and was approximately mesokurtic. Figure E.1 is a histogram of the composite scores. Finally, the average level of fluency in English was 1.8, which would correspond to "More fluent in English than in my first language." The distribution was positively skewed.

Frequencies for sex, ethnicity, and English fluency are reported in table C.19. 34% of respondents identified as male, 65% identified as female. One participant chose to not indicate their sex and was thus excluded from the analysis. 51% of respondents identified as Asian/Asian American/Pacific islanders, while 33.4% identified as European/American-/Caucasian. 8.6% identified as other and .5% did not report their ethnicity. For English fluency, 53% of participants said they were very fluent in English and that it is their first language. 22.3% reported being more fluent in English than in their first language. 13% reported being equally fluent in English and their first language, and 10.2% reported that they were less fluent in English than in their first language.

Descriptive statistics were then obtained for age and composite CES-D score by sex and are reported in table C.20. For males, the average age was 19.95 years, the maximum being 53 years of age and the standard deviation being 2.96. Females had similar results, the average age was 19.7 with a standard deviation of 2.81 and the same range. For the CES-D composite scores, scores ranged from 0-46 for males and the average was 14.21 with a 8.41 standard deviation. Females had a slightly larger range (0-51), the average composite was over 1 point higher (15.66) and the standard deviation was larger (9.44). Figure E.3 is a histogram for composite CES-D scores by sex.

Descriptive statistics by level of English fluency are provided in table C.21. For individuals who reported being "Very fluent in English; English is my first language," the average age was 19.66 years and the standard deviation was 2.85. Individuals who reported that they were "More fluent in English than in my first language" were 19.36 years of age, on average. Individuals who reported having the "Same fluency in English as my first language" had CES-D composite scores ranging from 0-48 and the average was 14.32 with a standard deviation of 9.14. For individuals who reported being "Less fluent in English than in my first language," the average age was somewhat older (21.44 years) with more variability (SD=3.49). For the CES-D, participants who were very fluent in English had an average

composite score of 14.32, with a standard deviation of 9.14. Individuals who were more fluent in English had slightly higher CES-D composite scores, an average of 15.33, with a standard deviation of 8.53. Individuals who were equally fluent in English and their first language had higher CES-D composite scores on average (16.11) and a standard deviation of 9.23. Individuals who were less fluent in English than in their first language had the highest average for CES-D composite scores at 18.04. There was also slightly more variability in these scores as indicated by a standard deviation of 9.49. Figure E.2 is a histogram of composite CES-D scores by level of English fluency.

4.3 Linear Mixed Models

A number of linear mixed effects models were tested for each feature. This section will describe the general process we used and how the models were defined.

All features were treated as continuous predictors, there were no factors entered in the model. English fluency was treated as a covariate. We made this choice because we thought there is clear rationale for this variable at least being at the ordinal scale level. Therefore, treating it as a categorical factor would not be appropriate. While this is a potential limitation of the project, we thought that the scale used was better considered as a quantitative variable than a categorical variable as someone who reported being less fluent in English than in their first language was almost certainly lower on the construct of English fluency than someone who was a monolingual speaker of English. Because a single individual chose "No answer" when responding to their sex, we did not include this as a level of the sex factor and this individual was excluded from subsequent analyses to ensure that the same data were included for each. Thus, all individuals included in the analysis reported their sex and English fluency.

To address issues of non-convergence, we took several approaches. To start, we limited the number of iterations that the software was given to come to a solution. Generally, the number of step halvings for each iteration was limited to 10, in some cases we increased the criterion to allow for 20 step-halvings. When a model did not converge, the iteration history was examined and either the syntax was edited to allow for more iterations, or a different covariance structure was used.

4.3.1 Model Fitting

For each domain (e.g., descriptive, morphological, syntactic), all the features were included in a single model, plus sex and English fluency, to determine a covariance structure to begin with. The choice of covariance structure was based on several considerations. Convergence was the first criterion, if the software did not find a solution for the model within a set number of iterations, step halvings, scoring steps, and singularity, additional covariance structures were then used. When convergence was achieved when using different covariance

structures for the same model (i.e., the same predictors and outcome variable), values for Akaike's Information Criterion (AIC) and Schwarz's Bayesian Criterion (BIC) were used to select the covariance structure.

Initially, certain covariance structures were considered to be more optimal from a theoretical perspective than others, which might be optimal from a computational perspective. For example, a covariance structure such as diagonal reduces computational complexity because fewer parameters are being estimated, thus models with this covariance structure are more likely to converge. However, this model does not make good sense theoretically. Thus, we used covariance structures such as autoregressive and Toeplitz when fitting models.

- A single feature model (i.e., the feature was included in the model as the only predictor) was obtained for each feature. Several covariance structures were tested for each feature and the best fitting-covariance structure was used for subsequent analyses. In some cases where a model did not converge, but the software reported that it would have a better fit (due to more parameters), the covariance structure which allowed for convergence was used. If the predictor was not found to be statistically significant, it was dropped from subsequent analyses.

- A three predictor model was obtained which included the feature, sex, and English fluency. Again, several covariance structures were tested for each model to find the best fitting solution. If the feature was not found to be statistically significant, no further analyses were run. If sex or English fluency were not found to be statistically significant, interaction models were not tested.

- Two four-predictor models were tested for each feature. These included the feature, sex, English fluency, and an interaction term. One model included a sex by feature interaction term while the other included a English fluency by feature interaction term. If convergence was not achieved for both interaction models, then no further higher order models were tested with the feature.

- A five-predictor model was then tested for some features. This model included a partial main effect for the feature, English fluency, and sex, a feature by English fluency interaction term, and a feature by sex interaction term.

- Finally, a small number of multiple feature models were tested in order to strengthen conclusions about these features. These models included features from multiple domains that appeared to be having large effects in the single predictor models. We started out with a model that contained only features as predictors, and then went on to add and remove certain features, as well as include Sex and English fluency.

To summarize, a considerable portion of this project was dedicated to model fitting and each set of predictors was tested using 2-6 covariance structures. To be sure, the

specific estimates were largely ignored until the best fitting models were selected to avoid any possibility for "p-hacking". In addition, as an exploratory study, the purpose was not to find support for a particular hypothesis. There are advantages to such a data-driven approach which prioritizes model fit over substantive theory testing. Namely, this approach simply tries to find models that fit the data well, rather than find models that fit a particular theory well. There are also disadvantages to this approach. This will be discussed further in the discussion chapter. The next section will further address how issues related to statistical significance were handled.

4.3.2 Type-I Error Control

To control for type I errors, we considered the domains separately as well as the categories of models (i.e., single predictor, three-predictor, interaction). We therefore had 5 "families" for each domain. Since a number of models were fit for each feature using different covariance structures (an average of 4 per model), this was taken into consideration. Our formula was thus:

$p = .05/(4 \times numberoffeaturesindomain)$

The per-test error rate used to determine signify statistical significance for a predictor is therefore different depending on the domain. The adjusted p-values will be included in each section.

While we were not mainly interested in significant results, due to the large number of tests, we used statistical significance as a way of excluding some variables from additional higher-order models.

4.4 Domain: Length

The length models included two features selected from the correlational analyses. The per test error rate used for each of the F-tests was .00625.

4.4.1 1-Predictor Models

Linear mixed effect models were specified to determine if the length of the item was predictive of the item responses. The length features chosen included the length in terms of the count of words as well as the average length of letters in each word in the item. For the length in terms of word count, a heterogeneous compound symmetric covariance structure was specified and the results of the tests of fixed effects showed $F_{(1,11170)}$ =56.31, $p < .006$. In particular, for each additional word, the item response increased by .005 points, indicating that longer items had higher endorsements of depressive symptomology, on average.

For the average number of letters per word in each item, a heterogeneous compound symmetric covariance structure was specified as well. The test for fixed effects indicated

$F_{(1,7001)}$ =279.27, $p < .006$. The estimates of fixed effects indicated that a one letter increase in the average number of letters per word in an item was associated with a .06 decrease in the item response, indicating that items with longer words were related to lower endorsement of depressive symptomology on average. The results for the coefficient estimates are included in D.1.

4.4.2 3-Predictor Models: Feature with English Fluency and Sex

Three predictor models were tested to determine if the length variables had a relationship with the responses when in the presence of individual characteristics such as English Fluency and Sex. For the number of words variable, models which specified a heterogeneous compound symmetric, compound symmetric, heterogeneous Toeplitz, and Toeplitz covariance structure did not achieve convergence. A heterogeneous autoregressive-1 covariance structure converged to provide estimates. Tests of fixed effects indicated $F_{(1,10100)}$ =76.78, $p < .006$ for the number of words, $F_{(1,11229)}$ =58.79, $p < .006$ for sex, and $F_{(1,11228)}$ =109.20, $p < .006$ for English Fluency. The estimates of fixed effects indicated that for each additional word in the item, the response increased by .01, controlling for other predictors. Males responses were .08 points lower than females, and for a 1 unit increase in English fluency, responses were .05 points higher on average, when controlling for other predictors.

For the feature of mean length of words in the item in terms of letters, a Toeplitz covariance structure was specified and achieved convergence. The tests of fixed effects were as follows: for average word length, $F_{(1,21095)}$=641.91, $p < .006$, for sex, $F_{(1,2166)}$ =11.3, $p < .006$, and for English Fluency, $F_{(1,2166)}$=36.70, $p < .006$. The estimates of the effects indicated that, when holding other predictors constant, as the average word length increased by 1 letter, the response on the CES-D decreased by .12, Males responses were .07 points lower than females, and for a 1 unit increase in English Fluency, the CES-D response was .06 points higher. The results for the coefficient estimates are included in D.2.

4.4.3 4-Predictor Models: Interaction Models

English Fluency by Feature Interaction

Four predictor models were tested to determine if the length features interacted with level of English fluency. For item length in terms of number of words, a Toeplitz covariance structure was specified. Tests of fixed effects were as follows: for number of words, $F_{(1,5460)}$ =75.24, $p < .006$, for sex, $F_{(1,2167)}$ =11.11, $p < .006$, for English fluency, $F_{(1,5454)}$ =31.05, $p < .006$ and for the interaction between number of words and English fluency, $F_{(1,5454)}$=.37, $p = .54$. The estimates of fixed effects indicated that for every 1 word increase in the number of words, the average CES-D response increased by .02, males' responses were .07 points lower than females', and for every 1 unit increase in English fluency, the average response increased by

.06 points. Of course, the interpretation of main effects in the presence of interactions is tenuous. The interaction was not statistically significant in this case.

For the average word length in terms of letters, a heterogeneous covariance structure was specified. The results of fixed effects were as follows: for average word length, $F_{(1,6924)} = 24.17, p < .006$, for sex, $F_{(1,2277)} = 52.15, p < .006$, for English fluency, $F_{(1,7169)} = 19.70, p < .006$, and for the interaction between English fluency and average word length, $F_{(1,6923)} = 15.77, p < .006$. The estimates for fixed effects indicated that, controlling for the other predictors, for an increase in the average word length by one letter, responses decreased by .04, males' responses were .18 points lower than females, for every 1 unit increase on English fluency, the CES-D responses increased by .09 points, and for the interaction between English fluency and average word length, the estimate was -.01, indicating that responses for individuals who were less fluent in English increased as the average word length decreased at a greater rate than the responses for individuals who were more fluent in English. The results for the coefficient estimates are included in D.3. The statistically significant interaction effect suggests that the effect of word length is stronger at lower levels of English fluency (i.e., higher levels of non-English dominance). In other words, it seems that longer words length slightly suppresses the effect of non-English dominance.

Sex by Feature Interaction

Four predictor models were tested to determine if the effects of length features interacted with the sex of the participants. For item length (in terms of number of words), a compound symmetric covariance structure was specified. Results of the tests of fixed effects were as follows: for item length, $F_{(1,21072)} = 278.55, p < .006$, for sex, $F_{(1,4272)} = 11.05, p < .006$, for English fluency, $F_{(1,2169)} = 35.78, p < .006$, and for the interaction between sex and item length, $F_{(1,41072)}$=.50, $p = .48$. Because the interaction term was not statistically significant in this model, we will not interpret the estimates of the fixed effects.

For the average word length, a heterogeneous compound symmetric covariance structure was specified. The tests of fixed effects were as follows: for the average word length, $F_{(1,6949)} = 228.05, p < .006$, for sex, $F_{(1,7057)} = 42.76, p < .006$, for English fluency, $F_{(1,2277)} = 4.04, p = .045$, and for the interaction between sex and average word length, $F_{(1,6949)}$=7.53, $p < .006$. The coefficient estimates suggest that, controlling for other variables, a one letter increase in word length corresponds to a .071 decrease in the CES-D response, a one unit increase in English fluency (non-English dominance) corresponds to a .023 increase in the CES-D response, male responses were .28 lower than female responses, and the interaction between sex and item length was .022. The results for the coefficient estimates are included in D.4. These results suggest that longer items dampen the effect of being male to a small extent. In other words, if we look at males, responses to a item that has longer words are closer to the female responses for those items than the responses to an item with shorter words.

4.4.4 5-Predictor Interaction Model

A five predictor model was obtained for the average word length in terms of letters. This model included a partial main effect for the feature, English fluency, and sex, and an interaction term for feature by sex and feature by English fluency. A heterogeneous compound symmetric covariance structure was specified. Results of the tests of fixed effects were as follows: for item length, $F_{(1,6950)}$=19.2, $p < .006$, for sex, $F_{(1,7078)}$=42.13, $p < .006$, for English fluency, $F_{(1,7084)}$=19.34, $p < .006$, $F_{(1,6948)}$=7.33, $p < .006$ for the interaction between sex and item length, and $F_{(1,6950)}$=15.58, $p < .006$. The results for the coefficient estimates are included in D.5.

4.4.5 Summary

For number of words as well as average word length in letters, the coefficient estimates in the three-predictor models were larger than in the 1-predictor models. For the average word length in letters, the effect was -.118. This means that the item with the highest average word length (6.67) was predicted to be nearly half a point (.472) lower on average than the item with the lowest word length. There was a small interaction effect between English fluency and average number of words. This suggested that for an individual who was non-English dominant, longer words on average resulted in a lower endorsement of depressive symptomology than short words. There was an interaction between sex and average word length such that the difference between males and females for items with shorter words was bigger than the difference between males and females for items with longer words.

4.5 Domain: Morphology I-Parts of Speech

There were several parts of speech variables that were examined for the morphological domain. These included nouns, verbs, adverbs, pronouns, adjectives, auxiliary verbs, copular verbs, and prepositions. All were included as proportions (i.e., the count of the part of speech divided by the total number of words) rather than counts. The per-test error rate was set to .0016.

4.5.1 1-Predictor Models

A single predictor model was fit for each feature and multiple covariance structures were specified. For the proportion of nouns, a heterogeneous Toeplitz covariance structure was selected among other potential options. Results of these analyses are included in table D.6. The results of fixed effects were as follows: $F_{(1,8357)} = 810.30, p < .0016$. Estimates of fixed effects indicated that for .1 increase in the proportion of nouns, the average response decreased by .077 points. That is, a higher proportion of nouns was associated with less endorsement of depressive symptomology. For the proportion of main verbs, a heterogeneous

compound symmetric covariance structure was specified. The results of the tests of fixed effects were as follows: $F_{(1,25662)} = 54.54, p < .0016$. Estimates of fixed effects indicated that for a .1 increase in the proportion of verbs, the CES-D response increased .019 points on average. That is, a higher proportion of verbs was associated with higher endorsement of depressive symptomology. For adverbs, a heterogeneous compound symmetric covariance structure was selected. The results of the tests of fixed effects were as follows: $F_{(1,13303)} = 185.59, p < .0016$. For the estimates of fixed effects, a .1 increase in the proportion of adverbs resulted in a .059 increase in the CES-D response, indicating that a higher proportion of adverbs was associated with higher endorsement of depressive symptomology.

For proportion of pronouns, a heterogeneous compound symmetric covariance structure was also specified. The tests of fixed effects were as follows: $F_{(1,10901)} = 309.45, p < .0016$. The estimates of fixed effects indicated that a .1 increase in the proportion of pronouns was associated with a .062 increase in the CES-D response. For adjectives, an heterogeneous Toeplitz covariance structure was specified. Tests of fixed effects were as follows: $F_{(1,9489)} = 26.18, p < .0016$. The estimates of fixed effects were as follows: for each .1 increase in the proportion of adjectives, there was a .011 decrease in the CES-D response on average. For auxiliary verbs, a heterogeneous Toeplitz covariance structure was specified for the model. The tests of fixed effects were as follows: $F_{(1,7856)} = 6.57, p = .010$. The estimates of fixed effects indicated that a .1 increase in the proportion of auxiliary verbs resulted in a .015 decrease in the CES-D response, on average. Because these results were not statistically significant when correcting for multiple tests, auxiliary verbs were not included in subsequent models. For copular verbs, a heterogeneous Toeplitz covariance structure was specified for the model. The results of the fixed effects were as follows $F_{(1,7932)} = 59.88, p < .0016$. The estimates of fixed effects indicated that for every .1 increase in the proportion of copular verbs, the average CES-D response decreased by .021. Finally, a heterogeneous Toeplitz covariance structure was specified for the proportion of prepositions. The results for the tests of fixed effects were as follows: $F_{(1,7364.14)} = 220.41, p < .0016$. The estimates of fixed effects indicated that for a .1 increase in the proportion of prepositions, the average CES-D response increased by .075.

In summary, single predictor mixed-effects models indicated that higher proportions of verbs, adverbs, pronouns and prepositions were related to higher endorsement of depressive symptomology while higher proportions of nouns, adjectives, and copular verbs were related to lower endorsement of depressive symptomology.

4.5.2 3-Predictor Models: Feature with English Fluency and Sex

Three predictor models included the feature variables, English fluency, and sex as predictors. For nouns, a Toeplitz covariance structure was specified for this model. Results of these analyses are provided in table D.7. Tests of fixed effects were as follows: $F_{(1,8340)} = 314.95, p < .0016$ for proportion of nouns, $F_{(1,2164.31)} = 35.71, p < .0016$ for En-

glish fluency and $F_{(1,2164)} = 12.59, p < .0016$ for sex. The estimates of the fixed effects were as follows: for every .1 increase in proportion of nouns, while holding the other predictors constant, there was a .056 decrease in CES-D responses. For English fluency, each unit decrease in level of English fluency resulted in a .06 increase in CES-D response. That is, being less fluent in English was associated with higher CES-D responses, when controlling for sex and proportion of nouns. Males responses, on average were .07 points lower than females. For main verbs, a compound symmetric covariance structure was specified. Results of tests of fixed effects were as follows: $F_{(1,41072)} = 38.68, p < .0016$ for main verbs, $F_{(1,2169)} = 35.83, p < .0016$ for level of English fluency and $F_{(1,2169)} = 12.22, p < .0016$ for sex. The estimates of fixed effects were as follows: holding other predictors constant, for each additional .1 increase in the proportion of main verbs, the CES-D response increased by .018, a 1 unit increase in English fluency was associated with a .06 point increase in CES-D score, and male responses on average were .07 points lower than female responses.

For adverbs, a Toeplitz covariance structure was specified. The results from the tests of fixed effects were as follows: $F_{(1,14109)} = .363, p = .547$ for proportion of adverbs, $F_{(1,2165)} = 36.03, p < .0016$ for level of English fluency, and $F_{(1,2165)} = 12.174, p < .0016$ for sex. Because the results did not indicate that the proportion of adverbs had a large effect on CES-D scores when controlling for sex and English fluency, no further models containing adverbs will be discussed. For pronouns, a heterogeneous compound symmetric covariance structure obtained the best fit. The results of the tests of fixed effects were as follows: $F_{(1,10734)} = 320.65, p < .0016$ for proportion of pronouns, $F_{(1,2374)} = 4.19, p = .04$ for level of English fluency, and $F_{(1,2374)} = 55.88, p < .0016$ for sex. The estimates of fixed effects indicated that, holding other predictors constant, for a .1 increase in the proportion of pronouns, the CES-D score increased by .063 (i.e., more pronouns was associated with more endorsement of depressive symptomology on average), a one level increase in English fluency (i.e., an increase in non-English dominance) lead to a .02 point increase in the CES-D score, and male responses were .19 points lower on average than female responses.

For adjectives, a heterogeneous Toeplitz covariance structure was specified. Results of the tests of fixed effects were as follows: $F_{(1,9471)} = 22.03, p < .0016$ for proportion of adjectives, $F_{(1,2355)} = 2.82, p = .09$ for English fluency, and $F_{(1,2355)} = 48.61, p < .0016$ for sex. The estimates of fixed effects were as follows: for a .1 increase in the proportion of adjectives, the CES-D response was decreased by .01, the effect of English fluency was not statistically significant and will not be interpreted, and males scores on average were .18 points lower than females. For copular verbs, a Toeplitz covariance structure was specified. Results of the tests of fixed effects were as follows: $F_{(1,13160)} = 69.30, p < .0016$ for proportion of copular verbs, $F_{(1,2165)} = 35.8, p < .0016$ for English fluency, and $F_{(1,2165)} = 12.86, p < .0016$ for sex. The estimates of fixed effects were as follows: holding all other predictors constant, for a .1 increase in the proportion of copular verbs, the CES-D score decreased by .026, for a 1 unit increase in level of English fluency (i.e., non-English dominance), the CES-D score

increased by .06, and male responses were .07 points lower than female responses on average. For prepositions, a heterogeneous Toeplitz covariance structure was specified. Results of the tests of fixed effects were as follows: $F_{(1,7326)} = 217.36, p < .0016$ for proportion of prepositions, $F_{(1,2112)} = 4.98, p = .03$ for level of English fluency, and $F_{(1,2112)} = 45.91, p < .0016$ for sex. The estimates of fixed effects were as follows: holding all else constant, for a .1 increase in the proportion of prepositions, the average CES-D response increased by .075. The estimates for the level of English fluency were not statistically significant. For sex, male responses were .17 points lower than female responses on average.

4.5.3 4-Predictor Models: Interaction Models

English Fluency by Feature Interaction

For nouns, a heterogeneous compound symmetric covariance structure had the best fit. Results of these analyses are provided in table D.8. Tests of fixed effects were as follows: $F_{(1,25225)} = 116.44, p < .0016$ for proportion of nouns, $F_{(1,2177)} = 11.49, p < .0016$ for English fluency, $F_{(1,1965)} = 48.61, p < .0016$ for sex, and $F_{(1,25223)} = 6.11, p = .013$ for the interaction between the proportion of nouns and English fluency. For the proportion of verbs, a Toeplitz covariance structure was used. Results were as follows: $F_{(1,25245)} = 29.63, p < .0016$ for proportion of main verbs, $F_{(1,2853)} = 3.95, p = .05$ for English fluency, $F_{(1,2572)} = 54.42, p < .0016$ for sex, and $F_{(1,25240)} = 3.46, p = .06$ for the interaction between proportion of main verbs and English fluency. For the proportion of adjectives, a heterogeneous autoregressive-1 covariance structure was specified. Results were as follows: $F_{(1,26612)} = 64.32, p < .0016$ for the proportion of adjectives, $F_{(1,14816)} = 47.85, p < .0016$ for English fluency, $F_{(1,11113)} = 59.33, p < .0016$ for sex, and $F_{(1,26614)} = 1.85, p = .17$ for the interaction between proportion of adjectives and English fluency. For the proportion of copular verbs, a Toeplitz covariance structure was specified. Results were as follows: $F_{(1,24373)} = 15.88, p < .0016$ for the proportion of copular verbs, $F_{(1,2385)} = 31.2, p < .0016$ for English fluency, $F_{(1,2166)} = 11.51, p = .007$ for sex, and $F_{(1,24345)} = 1.93, p = .17$ for interaction between proportion of copular verbs and English fluency. The model for the proportion of prepositions specified a Toeplitz covariance structure. Results were as follows: $F_{(1,13637)} = 92.78, p < .0016$ for the proportion of prepositions, $F_{(1,2403)} = 36.42, p < .0016$ for English fluency, $F_{(1,2167)} = 11.24, p < .0016$ for sex, and $F_{(1,13616)} = .025, p = .88$ for the interaction between proportion of prepositions and English fluency.

Sex by Feature Interaction

For the proportion of nouns, a heterogeneous autoregressive-1 covariance structure was specified. Results of these analyses are provided in table D.9. Results were as follows: $F_{(1,40364)} = 120.07, p < .0016$ for the proportion of nouns, $F_{(1,11547)} = 122.66, p < .0016$ for English fluency, $F_{(1,25892)} = 24.31, p < .0016$ for sex, and $F_{(1,40364)} = .05, p = .82$

for the interaction between proportion of nouns and sex. The model for the proportion of main verbs used a heterogeneous compound symmetric covariance structure. Results were as follows: $F_{(1,14952)} = 32.9, p < .0016$ for the proportion of main verbs, $F_{(1,2166)} = 35.75, p < .0016$ for English fluency, $F_{(1,4110)} = 1.90, p = .17$ for sex, and $F_{(1,14952)} = 8.02, p = .005$ for the interaction between proportion of main verbs and sex. The model for the proportion of adjectives used a Toeplitz covariance structure. Results were as follows: $F_{(1,19358)} = 286.76, p < .0016$ for the proportion of adjectives, $F_{(1,2166)} = 37.49, p < .0016$ for English fluency, $F_{(1,2999)} = .79, p = .37$ for sex, and $F_{(1,19358)} = 31.44, p < .0016$ for the interaction between proportion of adjectives and sex. For the proportion of copular verbs, a heterogeneous autoregressive-1 covariance structure achieved convergence. Results were as follows: $F_{(1,9148)} = 142.24, p < .0016$ for the proportion of copular verbs, $F_{(1,11102)} = 107.27, p < .0016$ for English fluency, $F_{(1,13681)} = 80.15, p < .0016$ for sex, and $F_{(1,9148)} = 28.84, p < .0016$ for the interaction between copular verbs and sex. for the proportion of prepositions, a Toeplitz covariance structure was specified. Results were as follows: $F_{(1,13636)} = 348.61, p < .0016$ for the proportion of prepositions, $F_{(1,2167)} = 37.91, p < .0016$ for English fluency, $F_{(1,2403)} = 12.66, p < .0016$ for sex, and $F_{(1,13636)} = 1.68, p = .19$ for the interaction between proportion of prepositions and sex.

4.5.4 5-Predictor Models

Because none of the interaction effects were statistically significant in both English fluency and sex interaction models for the part of speech features, no 5-predictor models were tested.

4.5.5 Summary

In the single predictor models, nouns, adjectives, auxiliary verbs, and copular verbs were related to lower endorsement of depressive symptomology, while main verbs, adverbs, pronouns, and prepositions were related to greater endorsement of depressive symptomology. Nouns and prepositions had the largest effects, followed by pronouns and adverbs. However, these effects were still very small. For example, an item without any nouns was .25 points higher on average than an item where a third of the words were nouns, which was the maximum proportion of nouns observed in these items. If 10 items had the maximum proportion of nouns (.33), these results would predict that the composite score would be 2.5 points lower (i.e., less endorsement of depressive symptomology) than if all of the items had no nouns. When controlling for sex and English fluency, the effect would predict a composite score 1.8 points lower if 10 items had a .33 proportion of nouns. Thus, this effect seems to have potential to be somewhat meaningful.

The models that included interactions were interesting substantively. The model with noun by English fluency interactions suggested that the effect of proportion of nouns was stronger at higher levels of non-English dominance. That is, the interaction effect in an item

for which the proportion of nouns was .33 was -.02 points for those who were most dominant in English, while it was -.09 points lower for those who were least dominant in English. In other words, there were stronger effects for nouns when non-English dominance was high. For verbs, the interaction effect indicated that higher levels of non-English dominance counterbalanced the effect of main verbs. This is evident in the opposite direction of the interaction effect and the main effect for verbs.

For the sex by feature interactions, females had CES-D scores that were .06 points higher than males for items in which the proportion of verbs was .33, intensifying the positive effect of sex (for females). There was an interaction between adjectives and sex such that the difference in the average CES-D score between males and females was larger when there were higher proportions of adjectives compared to when there were lower proportions of adjectives.

4.6 Domain: Morphology II-Other Features

For other features related to morphology that were not parts of speech, a per-test error rate of .004 was set. Counts of bound morphemes and suffixes were included, as well as the proportion of content words. These features were chosen based on the results of the correlational analysis.

4.6.1 1-Predictor Models

For the count of bound morphemes, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,21150)} = 375.16, p < .004$ for the count of bound morphemes. The estimates of fixed effects were as follows: for each additional bound morpheme, the CES-D score decreased by .09. for the count of suffixes, a heterogeneous compound symmetric was also specified. Results were as follows: $F_{(1,13784)} = 46.49, p < .004$ for the count of suffixes. The estimates of fixed effects were as follows: for each additional suffix, the CES-D score decreased by .04. The model for the proportion of content words specified a heterogeneous compound symmetric covariance structure. Results were as follows: $F_{(1,13341)} = 224.26, p < .004$ for the proportion of content words. The estimates of fixed effects were as follows: for a .1 increase in the proportion of content words, the CES-D score decreased by .033. The coefficient estimates are included in D.10.

4.6.2 3-Predictor Models: Feature with English Fluency and Sex

The model for the count of bound morphemes specified a heterogeneous compound symmetric covariance structure. Results were as follows: $F_{(1,21070)} = 378.47, p < .004$ for the count of bound morphemes, $F_{(1,2199)} = 6.84, p = .009$ for English fluency, and $F_{(1,2200)} = 43.33, p < .004$ for sex. The estimates of fixed effects were as follows: holding all other

predictors constant, for each additional bound morpheme, the CES-D response decreased by .26, for an increase in the level of English fluency, the CES-D response decreases by .79., and male responses were lower than female responses by .33.

For the count of suffixes, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,13792)} = 82.31, p < .004$ for the count of suffixes, $F_{(1,2583)} = 2.78, p = .10$ for English fluency, and $F_{(1,2583)} = 46.18, p < .004$ for sex. The estimates of fixed effects were as follows: for each additional suffix, the CES-D score decreased by .04. The results were not statistically significant for English fluency, and male responses were lower than female responses by .17. Because English fluency was not statistically significant, no further tests were included where suffixes were a predictor.

For the proportion of content words, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,13407)} = 223.2, p < .004$ for proportion of content, $F_{(1,2326)} = 3.05, p = .08$ for English fluency, and $F_{(1,2326)} = 47.3, p < .004$ sex. The estimates of fixed effects were as follows: for a .1 increase in the proportion of content words, the CES-D score decreased by .033. The results were not statistically significant for English fluency, and males' responses were lower than females' by .17. Estimates for coefficients are included in D.11.

4.6.3 4-Predictor Models: Interaction Models

English Fluency by Feature Interaction

For the count of bound morphemes, a heterogeneous autoregressive covariance structure was specified. Results were as follows: $F_{(1,14815)} = 11.19, p = .001$ for bound morphemes, $F_{(1,11153)} = 59.77, p < .004$ for sex, $F_{(1,17499)} = 192.33, p < .004$ for English fluency, and $F_{(1,14816)} = 81.54, p < .004$ for the interaction between English fluency and bound morphemes. Estimates for coefficients are included in D.12.

Sex by Feature Interaction

The model for the count of bound morphemes specified a heterogeneous compound symmetric covariance structure. Results were as follows: $F_{(1,21081)} = 389.94, p < .004$ for the count of bound morphemes, $F_{(1,2207)} = 7.02, p = .008$ for English fluency, $F_{(1,2601)} = 13.85, p < .004$ for sex, and $F_{(1,21081)} = 16.15, p < .004$ for the interaction between the count of bound morphemes and sex. Estimates for coefficients are included in D.13.

4.6.4 5-Predictor Model

A model for the count of bound morphemes was obtained with an interaction effect for bound morphemes by sex and bound morphemes by English fluency. This model specified a heterogeneous compound symmetric covariance structure. Results were as follows:

$F_{(1,21097)} = 7.03, p = .008$ for the count of bound morphemes, $F_{(1,2613)} = 46.24, p < .004$ for English fluency, $F_{(1,2615)} = 13.10, p < .004$ for sex, $F_{(1,21091)} = 17.01, p < .004$ for the interaction between sex and bound morphemes, and $F_{(1,21111)} = 79.24, p < .004$ for the interaction between English fluency and bound morphemes. Estimates for coefficients are included in D.14. Estimates of fixed effects were as follows: holding all other predictors constant, for each additional bound morpheme, the CES-D response decreased by .005, for a one level increase in English fluency, the CES-D response increased by .089, males' responses were lower than females' by .10, the interaction between bound morphemes and sex was -.039 and the interaction between level of English fluency and bound morphemes was -.039. This suggests that the effect of bound morphemes was stronger for males than it was for females and the effect of bound morphemes was stronger for individuals who were non-dominant in English.

4.6.5 Summary

The effect of bound morphemes was quite large in the single predictor model, suggesting that an item with 3 bound morphemes received responses that were lower than an item with no bound morphemes by .27. Across 10 items, that is a -2.7 point difference. This effect intensified when an individual was higher on non-English dominance, or male.

4.7 Domain: Syntax

The features examined within the domain of syntax included ambiguity, negations, possessives, and relative clauses. The per-test error rate was set to .003.

4.7.1 1-Predictor Models

For ambiguity, a heterogeneous compound symmetry covariance structure was specified. Results were as follows: $F_{(1,21915)} = 108.54, p < .003$ for the ambiguity. The estimates of fixed effects were as follows: the presence of syntactic ambiguity results in responses that are lower by .07. For negations, a heterogeneous compound symmetry was specified. Results were as follows: $F_{(1,32260)} = 124.21, p < .003$ for presence of a negation. The estimates of fixed effects were as follows: on average, items containing a negation were lower than items which do not contain a negation by .08. For the presence of a relative clause, a heterogeneous compound symmetry was specified. Results were as follows: $F_{(1,6005)} = 1615.74, p < .003$ for presence of a relative clause. The estimates of fixed effects were as follows: on average, items containing a relative clause had higher responses than items which do not contain a relative clause by .42. For possessives, a Toeplitz covariance structure achieved convergence and had the best fit . Results were as follows: $F_{(1,15959)} = 38.71, p < .003$ for items containing a possessive. The estimates of fixed effects were as follows: on average, items containing

a possessive had higher responses than items that do not contain possessives by .05. The coefficient estimates are included in D.15.

4.7.2 3-Predictor Models: Feature with English Fluency and Sex

For ambiguity, a Toeplitz covariance structure was specified. Results were as follows: $F_{(1,15112)} = 316.44, p < .003$ for ambiguity. The estimates of fixed effects were as follows: on average, items containing a syntactic ambiguity had lower responses than items that were not coded as having an ambiguous structure by .14, male responses were lower than female responses by .07, and for an increase in the level of English fluency,the CES-D responses increased by .06.

For negations, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,31650)} = 132.68, p < .003$ for the presence of a negation, $F_{(1,2480)} = 4.73, p = .03$ for English fluency, and $F_{(1,2480)} = 53.01, p < .003$ for sex. The estimates of fixed effects were as follows: items containing a negation were on average, .08 points lower than items that did not contain a negation, for an decrease in the level of English fluency, responses increased by .03, and male responses were -.19 points lower than female responses.

For possessives, a Toeplitz covariance structure was specified. Results were as follows: $F_{(1,17699)} = 44.52, p < .003$ for the presence of a possessive, $F_{(1,2166)} = 36.95, p < .003$ for English fluency, and $F_{(1,2166)} = 11.37, p < .003$ for sex. The estimates of fixed effects were as follows: on average, items containing a possessive received higher responses than those that did not contain a possessive by .06 points, for an decrease in the level of English fluency, responses increased by .06, and male responses were .07 points lower than female responses on average.

For the presence of a relative clause, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,20223)} = 2944.1, p < .003$ for the presence of a relative clause, $F_{(1,2167)} = 38.43, p < .003$ for English fluency, and $F_{(1,2167)} = 10.96, p < .003$ for sex. The estimates of fixed effects were as follows: items containing a relative clause were .54 points higher on average than items that did not contain a relative clause, for an increase in the level of English fluency, the CES-D response increased by .06, and male responses were .07 points lower than female responses on average. Coefficient estimates are included in D.16.

4.7.3 4-Predictor Models: Interaction Models

English Fluency by Feature Interaction

For ambiguity, a compound symmetric covariance structure was specified. Results were as follows: $F_{(1,41071)} = 35.13, p < .003$ for ambiguity, $F_{(1,2436)} = 39, p < .003$ for English

fluency, $F_{(1,2169)} = 12.22, p < .003$ for sex, and $F_{(1,41071)} = 3.26, p = .07$ for the interaction between ambiguity and English fluency.

For relative clauses, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,5938)} = 480.81, p < .003$ for the presence of a relative clause, $F_{(1,1271)} = 5.87, p = .016$ for English fluency, $F_{(1,1298)} = 39.47, p < .003$ for sex, and $F_{(1,5936)} = 6.11, p = .013$ for the interaction between relative clause and English fluency.

For possessives, a heterogeneous compound symmetry covariance structure was selected. Results were as follows: $F_{(1,15243)} = .307, p = .58$ for the presence of a possessive, $F_{(1,2656)} = 3.3, p = .07$ for English fluency, $F_{(1,2568)} = 52.41, p < .003$ for sex, and $F_{(1,15244)} = 9.25, p < .003$ for the interaction between possessives and English fluency. Coefficient estimates are included in D.17.

Sex by Feature Interaction

For ambiguity, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,22476)} = 123.86, p < .003$ for the presence of ambiguity, $F_{(1,2603)} = 2.83, p = .092$ for English fluency, $F_{(1,2690)} = 36.59, p < .003$ for sex, and $F_{(1,22476)} = 16.8, p < .003$ for the interaction between ambiguity and sex. For relative clauses, a Toeplitz covariance structure was specified. Results were as follows: $F_{(1,202300)} = 2653, p < .003$ for the presence of a relative clause, $F_{(1,2167)} = 38.43, p < .003$ for English fluency, $F_{(1,2270)} = 10.75, p < .003$ for sex, and $F_{(1,20230)} = .002, p = .96$ for the interaction between relative clauses and sex. For possessives, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,15667)} = 23.0, p < .003$ for the presence of a possessive, $F_{(1,2686)} = 2.32, p = .13$ for English fluency, $F_{(1,2596)} = 58.85, p < .003$ for sex, and $F_{(1,15667)} = 14.68, p < .003$ for the interaction between possessives and sex. Coefficient estimates are included in D.18.

4.7.4 5-Predictor Model

A model which included possessives, sex, English fluency, and a feature by sex and feature by English fluency interaction term was tested. A heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,15451)} = .004, p = .948$ for the presence of a possessive, $F_{(1,2666)} = 3.45, p = .063$ for English fluency, $F_{(1,2667)} = 59.10, p < .003$ for sex, and $F_{(1,15461)} = 14.26, p < .003$ for the interaction between possessives and sex, and $F_{(1,15448)} = 8.84, p = .003$ for the interaction between English fluency and possessives. Coefficient estimates are included in D.19.

4.7.5 Summary

For single predictor models, ambiguity, negations, and possessives all had negative effects on the CES-D response, while relative clauses had a positive effect. Relative clauses had a

particularly large effect that was even larger when controlling for individual characteristics. The interaction effect between relative clauses and English fluency suggested that the effect of English fluency was cancelled out when a relative clause was present as the coefficients had the same absolute value but opposite signs. The feature of whether an item contained a possessive was found to have statistically significant interactions with both sex and English fluency. However, like was seen with relative clauses, the magnitude of the interaction effect between possessives and English fluency is equal and opposite to the magnitude of English language fluency.

4.8 Domain: Semantics I - Manually Coded Features

Models were obtained for features within the domain of semantics. These features were those selected which were manually coded, or one of the CES-D subscales. The adjusted per-test error rate was .00125.

4.8.1 1-Predictor Models

For manually coded specificity, a heterogeneous compound symmetric covariance structure was specified. Coefficient estimates are provided in table D.20. Results were as follows: $F_{(1,14128)} = 30.93, p < .001$. The estimates of fixed effects were as follows: on average, items which were coded as being specific had lower responses than items that were coded as being general by .04. For manually coded concreteness, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,20502)} = 50.4, p < .001$ The estimates of fixed effects were as follows: on average, items that were coded as being concrete had responses that were lower than items that were coded as being non-concrete by .05. For manually coded cultural loading, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,33086)} = 1589.25, p < .001$ The estimates of fixed effects were as follows: on average, items which were coded as containing cultural loading received lower responses than items that were coded as being non-culturally loaded by .27. For manually coded gender (i.e., items that might have different responses based on the gender of the individual), a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,25851)} = 59.67, p < .001$ The estimates of fixed effects were as follows: on average, items that were coded as having features that might result in differential responses in respect to gender received lower responses than items that were coded as not having a gender component by .05.

For items that were on the positive affect subscale (and were thus reverse-scored), a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,9224)} = 694.66, p < .001$ The estimates of fixed effects were as follows: on average, items that were on the positive affect subscale had responses that were higher than items not on this subscale by .23. For items that were on the somatic subscale, a hetero-

geneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,14913)} = 976.83, p < .001$ The estimates of fixed effects were as follows: on average, items on the somatic subscale had higher responses than items that were not on the subscale by .23. For items that were on the depressed affect subscale, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,29502)} = 559.41, p < .001$ The estimates of fixed effects were as follows: on average, items that were on the depressed affect subscale had responses that were lower than items not on this subscale by .16. For items on the interpersonal problems subscale, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,11904)} = 422.36, p < .001$ The estimates of fixed effects were as follows: on average, items that were on the interpersonal problems subscale had responses that were lower than items not on this subscale by .2. For items that were coded as referring to others, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,21161)} = 182.99, p < .001$ The estimates of fixed effects were as follows: on average, items that referred to other people had responses that were higher than items that did not refer to other people by .11.

For items that referred to a behavior (rather than a feeling), a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,16420)} = 4.02, p = .04$ The estimates of fixed effects were as follows: Because the results were not statistically significant, we did not interpret the results here and this feature will not appear in subsequent analyses.

4.8.2 3-Predictor Models: Feature with English Fluency and Sex

For manually coded specificity, a heterogeneous compound symmetric covariance structure was specified. Coefficient estimates are available in D.21. Results were as follows: $F_{(1,14381)} = 28.90, p < .001$ for items coded as being specific, $F_{(1,2582)} = 2.68, p = .102$ for English fluency, and $F_{(1,2582)} = 47.37, p < .001$ for sex. The estimates of fixed effects were as follows: items that were coded as being specific received lower responses than items that were coded as being general by .04, because the effect for English fluency was not statistically significant, it will not be interpreted here, males' responses were lower than females' by .18. For items coded as being concrete, a heterogeneous compound symmetric covariance structure was specified.

Results were as follows: $F_{(1,20091)} = 47.86, p < .001$ for items coded as being concrete, $F_{(1,2495)} = 3.1, p = .08$ for English fluency, and $F_{(1,2495)} = 46.23, p < .001$ for sex. The estimates of fixed effects were as follows: items that were coded as being concrete received lower responses than items that were coded as being non-concrete by .05, because the effect for English fluency was not statistically significant, it will not be interpreted here, and male responses were lower than female responses by .17. For items that were culturally loaded, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,33080)} = 1596.23, p < .001$ for culturally loaded items, $F_{(1,1978)} = 11.22, p < .001$

for English fluency, and $F_{(1,1978)} = 39.49, p < .001$ for sex. The estimates of fixed effects were as follows: items that were culturally loaded had responses that were lower than non-culturally loaded items by .24, because the effect for English fluency was not statistically significant, it will not be interpreted here, and male responses were lower than female responses by .13.

For items coded as having some gender-related component, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,26774)} = 50.78, p < .001$ for gender coding, $F_{(1,2587)} = 2.48, p = .12$ for English fluency, and $F_{(1,2588)} = 42.67, p < .001$ for sex. The estimates of fixed effects were as follows: items that were coded as having a gender-related component had responses that were lower than other items by .05, because the effect for English fluency was not statistically significant, it will not be interpreted here, and male responses were lower than female responses by .17. For items that were on the positive affect subscale, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,9102)} = 705.28, p < .001$ for positive affect, $F_{(1,1909)} = 14.41, p < .001$ for English fluency, and $F_{(1,1910)} = 36.57, p < .001$ for sex. The estimates of fixed effects were as follows: items that were on the positive affect subscale had responses that were higher than other items by .23, as the level of English fluency increased, the responses were higher by .04, and males' responses were lower than females' by .14.

For items that were on the somatic symptoms subscale, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,14697)} = 969.1, p < .001$ for somatic symptoms, $F_{(1,1507)} = 3.93, p = .05$ for English fluency, and $F_{(1,1507)} = 36.31, p < .001$ for sex. The estimates of fixed effects were as follows: items on the somatic subscale received responses that were higher than other items by .23, the coefficient for English fluency was not statistically significant and will not be interpreted, and males' responses were lower than females' by .14. For items that were on the depressed affect subscale, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,28673)} = 533.91, p < .001$ for depressed affect, $F_{(1,1674)} = 4.74, p = .03$ for English fluency, and $F_{(1,1674)} = 32.56, p < .001$ for sex. The estimates of fixed effects were as follows: items on the depressed affect subscale had responses that were lower than other items by .16, the coefficient for English fluency was not statistically significant and will not be interpreted, and male responses were lower than female responses by .14.

For items that were on the interpersonal problems subscale, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,11594)} = 403.5, p < .001$ for interpersonal problems, $F_{(1,1176)} = 7.07, p = .008$ for English fluency, and $F_{(1,1176)} = 49.2, p < .001$ for sex. The estimates of fixed effects were as follows: items on the interpersonal problems subscale received responses that were lower than other items by .19, the coefficient for English fluency was not statistically significant and will not be interpreted, and male responses were lower than females by .17. For items that referred to other people, a heterogeneous compound symmetric covariance structure was specified. Results

were as follows: $F_{(1,20306)} = 189.21, p < .001$ for referring to others, $F_{(1,2375)} = 3.28, p = .07$ for English fluency, and $F_{(1,2375)} = 54.21, p < .001$ for sex. The estimates of fixed effects were as follows: items that referred to others had responses that were lower than other items by .11, the coefficient for English fluency was not statistically significant and will not be interpreted, male responses were lower than female responses by .19.

4.8.3 4-Predictor Models: Interaction Models

English Fluency by Feature Interaction

The model for positive affect specified a heterogeneous compound symmetric covariance structure. Coefficient estimates are provided in D.22. Results were as follows: $F_{(1,9476)} = 1.93, p = .165$ for positive affect, $F_{(1,1912)} = 15.54, p < .001$ for English fluency, $F_{(1,1915)} = 36.27, p < .001$ for sex, and $F_{(1,9474)} = 192.49, p < .001$ for the interaction between English fluency and positive affect.

The model for interpersonal problems specified a heterogeneous compound symmetric covariance structure. Results were as follows: $F_{(1,11709)} = 24.788, p < .001$ for interpersonal problems, $F_{(1,1180)} = 19.21, p < .001$ for English fluency, $F_{(1,1150)} = 47.73, p < .001$ for sex, and $F_{(1,11710)} = 32.33, p < .001$ for the interaction between interpersonal problems and English Fluency. The model for cultural loading specified a heterogeneous compound symmetric covariance structure. Results were as follows: $F_{(1,33081)} = 1464.35, p < .001$ for culturally loaded items, $F_{(1,1979)} = 11.22, p < .001$ for English fluency, $F_{(1,2084)} = 26.02, p < .001$ for sex, and $F_{(1,33081)} = 1.01, p = .31$ for the interaction between sex and cultural loading.

Sex by Feature Interaction

The model for positive affect specified a Compound symmetric covariance structure was specified. Results for the coefficient estimates are available in D.23. Results were as follows: $F_{(1,41071)} = 437.83, p < .001$ for positive affect, $F_{(1,2169)} = 35.81, p < .001$ for English fluency, $F_{(1,2322)} = 15.95, p < .001$ for sex, and $F_{(1,41071)} = 9.22, p < .001$ for the interaction between sex and positive affect.

The model for interpersonal problems specified a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,11903)} = 286.83, p < .001$ for interpersonal problems, $F_{(1,1350)} = 6.2, p = 013$ for English fluency, $F_{(1,1358)} = 73.43, p < .001$ for sex, and $F_{(1,11903)} = 29.53, p < .001$ for the interaction between interpersonal problems and sex.

The model for culture specified a compound symmetric covariance structure was specified. Results were as follows: $F_{(1,41071)} = 422, p < .001$ for culturally loaded items, $F_{(1,2680)} = 18, p < .001$ for English fluency, $F_{(1,2169)} = 12.22, p < .001$ for sex, and $F_{(1,41071)} = 20.41, p < .001$ for the interaction between cultural loading and sex.

4.8.4 5-Predictor Model

A five-predictor model was obtained for the interpersonal problems subscale. Results were as follows: $F_{(1,12013)} = 15.09, p < .001$ for interpersonal problems, $F_{(1,1337)} = 17.83, p = .013$ for English fluency, $F_{(1,1338)} = 71.45, p < .001$ for sex, $F_{(1,12008)} = 28.82, p < .001$ for the interaction between interpersonal problems and sex, $F_{(1,12021)} = 31.64, p < .001$ for the interaction between interpersonal problems and English Fluency. Coefficient estimates are included in D.24.

4.8.5 Summary

Cultural loading had a large negative effect on the responses, while items on the somatic subscale and the positive affect subscale were associated with greater endorsement of depressive symptomology on average. Referring to others and interpersonal problems were both related to lower endorsement of depressive symptomology. Culture, positive affect, and interpersonal problems all had statistically significant interaction effects with English fluency. For cultural loading, this effect suggested that for culturally loaded items, the positive effect of English fluency (non-English dominance) was strengthened. In other words, the negative effect of cultural loading was not as strong at higher levels of English fluency. The interaction between positive affect and English fluency suggested that the effect of positive affect was strongest for individuals who were less fluent in English. Finally, the interaction effect for the interpersonal problems subscale cancelled out the effect of English fluency, as the coefficients were the same magnitude but in opposite directions.

The interpersonal problems subscale showed a statistically significant interaction effect with sex. The main effects for interpersonal problems and sex were similar magnitudes and in a negative direction, whereas the interaction effect was about half the magnitude of either effect and in the opposite direction.

4.9 Domain: Semantics II - Coh-Metrix Features

4.9.1 1-Predictor Models

Results for the 1-predictor models can be found in table D.25. For the measure of polysemy based on FrameNet (manually calculated), a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,16663)} = 171.91, p < .001$. The estimates of fixed effects were as follows: on average, as an item's average number of meanings (polysemy) increased by 1, the CES-D response decreased by .02. For the measure of polysemy based on WordNet (manually calculated), a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,7852)} = 18.36, p < .001$. The estimates of fixed effects were as follows: on average, as an item's average number of meanings (polysemy) increased by 1, the CES-D response increased by .0036.

For the measure of word frequency for all words, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,18277)} = 18.93, p < .001$. For the measure of word frequency for all words, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,18277)} = 18.93, p < .001$. The estimates of fixed effects were as follows: on average, as the average frequency of content words in an item increased by 1, the CES-D response increased by .027. For the familiarity rating, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,13884)} = 63.41, p < .001$. The estimates of fixed effects were as follows: on average, as the average familiarity rating for the item increased by 1, the CES-D response increased by .0016.

For the average concreteness of words in an item, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,14500)} = 255.52, p < .001$. The estimates of fixed effects were as follows: on average, for a 1 unit increase in concreteness, the CES-D response decreased by .0014. For the average imageability of words in an item, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,17466)} = 222.42, p < .001$. The estimates of fixed effects were as follows: on average, for a 1 unit increase in imageability, the CES-D response decreased by .001.

For the average meaningfulness, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,14653)} = 19.49, p < .001$. The estimates of fixed effects were as follows: on average, for a 1 unit increase in meaningfulness, the CES-D response increased by .0002. For the polysemy measure from Coh-Metrix, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,9179)} = 5.02, p = .025$. The estimates of fixed effects were as follows: The results for polysemy will not be interpreted because the effect was not statistically significant. For hypernymy of nouns, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,26022)} = 329.02, p < .001$. The estimates of fixed effects were as follows: on average, for a 1 unit increase in hypernymy for nouns, the CES-D response decreased by .019. For hypernymy of verbs, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,4522)} = 178.41, p < .001$. The estimates of fixed effects were as follows: on average, for a 1 unit increase in hypernymy for verbs, the CES-D response increased by .06.

4.9.2 3-Predictor Models: Feature with English Fluency and Sex

Results for the coefficient estimates are available in table D.26. For polysemy that was manually coded using FrameNet, a Toeplitz covariance structure was specified. Results were as follows: $F_{(1,11819)} = 100.29, p < .001$ for polysemy, $F_{(1,2166)} = 37.02, p < .001$ for English fluency, and $F_{(1,2167)} = 11.40, p < .001$ for sex. The estimates of fixed effects were as follows: for a 1 unit increase in polysemy, the CES-D response decreased by .02, for a 1

unit increase in English fluency,the CES-D response increased by .06, and male responses were lower than female responses by .07.

For imageability, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,17705)} = 227.08, p < .001$ for imageability, $F_{(1,2323)} = 1.78, p = .182$ for English fluency, and $F_{(1,2323)} = 56.04, p < .001$ for sex. The estimates of fixed effects were as follows: for a 1 unit increase in imageability, the CES-D response decreased by .001, the results for English fluency were not statistically significant and will not be interpreted, and male responses were lower than female responses by .19.

For polysemy which was manually coded based on WordNet, a heterogeneous autoregressive-1 covariance structure was specified. Results were as follows: $F_{(1,6351)} = 118.04, p < .001$ for polysemy, $F_{(1,11181)} = 107.10, p < .001$ for English fluency, and $F_{(1,11182)} = 56.56, p < .001$ for sex. The estimates of fixed effects were as follows: for a 1 unit increase in polysemy, the CES-D response increased by .01, for a 1 unit increase in English fluency,the CES-D response increased by .051, and male responses were lower than female responses by .08. For frequency, a heterogeneous autoregressive-1 covariance structure was specified. Results were as follows: $F_{(1,12169)} = 69.48, p < .001$ for frequency, $F_{(1,11216)} = 110.08, p < .001$ for English fluency, and $F_{(1,11218)} = 58.63, p < .001$ for sex. The estimates of fixed effects were as follows: for a 1 unit increase in frequency, the CES-D response increased by .06, for a 1 unit increase in English fluency,the CES-D response increased by .051, and males' responses were lower than females' by .08.

For concreteness, a Toeplitz covariance structure was specified. Results were as follows: $F_{(1,11434)} = 1461.61, p < .001$ for concreteness, $F_{(1,2167)} = 36.92, p < .001$ for English fluency, and $F_{(1,2167)} = 11.49, p < .001$ for sex. The estimates of fixed effects were as follows: for a 1 unit increase in concreteness, the CES-D response decreased by .004, for a 1 unit increase in English fluency, the CES-D response increased by .06, and males' responses were lower than females' by .07. For hypernymy of nouns, a Toeplitz covariance structure was specified. Results were as follows: $F_{(1,13999)} = 9.83, p = .0017$ for hypernymy, $F_{(1,2166)} = 36.66, p < .001$ for English fluency, and $F_{(1,2166)} = 11.40, p < .001$ for sex. The estimates of fixed effects were as follows: for a 1 unit increase in hypernymy, the CES-D response increased by .004, for a 1 unit increase in English fluency, the CES-D response increased by .06, and male responses were lower than female responses by .07.

For meaningfulness, a heterogeneous autoregressive-1 covariance structure was specified. Results were as follows: $F_{(1,12413)} = 4.92, p = .027$ for meaningfulness, $F_{(1,11204)} = 111.13, p < .001$ for English fluency, and $F_{(1,11206)} = 55.77, p < .001$ for sex. The estimates of fixed effects were as follows: because the results for meaningfulness were not statistically significant, they will not be interpreted, for a 1 unit increase in English fluency, the CES-D response increased by .052, and male responses were lower than female responses by .08. For hypernymy of verbs, a heterogeneous compound symmetric covariance structure was specified. Results were as follows: $F_{(1,4531)} = 174.60, p < .001$ for hypernymy,

$F_{(1,2360)} = 2.57, p = .12$ for English fluency, and $F_{(1,2360)} = 47.00, p < .001$ for sex. The estimates of fixed effects were as follows: for a 1 unit increase in hypernymy, the CES-D response increased by .06, the results for English fluency were not statistically significant and will not be interpreted, and males' responses were lower than females' by .17.

4.9.3 4-Predictor Models: Interaction Models

English Fluency by Feature Interaction

For polysemy which was coded manually based on WordNet, a heterogeneous autoregressive-1 covariance structure was specified. Coefficient estimates are provided in table D.27. Results were as follows: $F_{(1,6315)} = 29.21, p < .001$ for polysemy, $F_{(1,16731)} = 30.89, p < .001$ for English fluency, $F_{(1,11182)} = 56.56, p < .001$ for sex, and $F_{(1,6321)} = .003, p = .96$ for the interaction between English fluency and polysemy. For polysemy, which was coded manually using FrameNet, a Toeplitz covariance structure was specified. Results were as follows: $F_{(1,11817)} = 14.42, p < .001$ for polysemy, $F_{(1,4708)} = 32.58, p < .001$ for English fluency, $F_{(1,2167)} = 11.401, p = .007$ for sex, and $F_{(1,11818)} = 1.74, p = .19$ for the interaction between polysemy and English fluency.

For frequency, a heterogeneous autoregressive-1 covariance structure was specified. Results were as follows: $F_{(1,12133)} = 7.49, p = .006$ for frequency, $F_{(1,10576)} = 1.06, p = .303$ for English fluency, $F_{(1,11217)} = 58.70, p < .001$ for sex, and $F_{(1,12128)} = 2.47, p = .116$ for the interaction between English fluency and frequency. For familiarity, a heterogeneous autoregressive-1 covariance structure was specified. Results were as follows: $F_{(1,11703)} = 21.48, p < .001$ for familiarity, $F_{(1,11574)} = .89, p = .344$ for English fluency, $F_{(1,11147)} = 61.70, p < .001$ for sex, and $F_{(1,11704)} = 1.74, p = .188$ for the interaction between English fluency and familiarity.

For concreteness, a heterogeneous autoregressive-1 covariance structure was specified. Results were as follows: $F_{(1,7818)} = 81.89, p < .001$ for concreteness, $F_{(1,10101)} = 48.10, p < .001$ for English fluency, $F_{(1,11283)} = 52.92, p < .001$ for sex, and $F_{(1,7823)} = 29.14, p < .001$ for the interaction between English fluency and concreteness. Coefficient estimates were as follows: holding other predictors constant, for a 100 point increase in concreteness, the CES-D response decreased by .2, male responses were lower than female responses by .08, a one unit increase in the level of English fluency was associated with a .22 increase in the CES-D response, and the interaction between English fluency and concreteness was -.001. This suggests that The effect of concreteness was stronger at higher levels of English fluency (i.e., non-English dominance). For hypernymy of nouns, a heterogeneous autoregressive-1 covariance structure was specified. Results were as follows: $F_{(1,24910)} = 25.64, p < .001$ for hypernymy for nouns, $F_{(1,13813)} = 107.21, p < .001$ for English fluency, $F_{(1,11153)} = 56.53, p < .001$ for sex, and $F_{(1,24913)} = 16.65, p < .001$ for the interaction between English fluency and imageability. Estimates of coefficients were as follows: holding other predictors constant, a one

unit increase in hypernymy of nouns (i.e., an item with more specific nouns) was associated with a .013 decrease in the CES-D response, male responses were .08 lower than female responses, a one unit increase in English fluency (i.e., increase in non-English dominance) was associated with a .07 increase in CES-D response, and the interaction between English fluency and hypernymy was .005. This suggests that the effect of hypernymy was stronger when non-English dominance was higher.

Sex by Feature Interaction

For polysemy, which was coded using WordNet, a heterogeneous autoregressive-1 covariance structure was specified. Coefficient estimates are provided in table D.28. Results were as follows: $F_{(1,6360)} = 139.12, p < .001$ for polysemy, $F_{(1,11179)} = 107.24, p < .001$ for English fluency, $F_{(1,16790)} = 65.95, p < .001$ for sex, and $F_{(1,6360)} = 23.55, p < .001$ for the interaction between sex and polysemy. For frequency, a heterogeneous autoregressive-1 covariance structure was specified. Results were as follows: $F_{(1,11884)} = 118.55, p < .001$ for frequency, $F_{(1,11199)} = 109.26, p < .001$ for English fluency, $F_{(1,10343)} = 124.49, p < .001$ for sex, and $F_{(1,11884)} = 91.22, p < .001$ for the interaction between sex and frequency.

For polysemy, which was coded using FrameNet, a Toeplitz covariance structure was specified. Results were as follows: $F_{(1,41070)} = 106.78, p < .001$ for polysemy, $F_{(1,2169)} = 35.8, p < .001$ for English fluency, $F_{(1,5175)} = .93, p = .33$ for sex, and $F_{(1,41070.33)} = 39.63, p < .001$ for the interaction between sex and polysemy. For hypernymy of nouns, a diagonal covariance structure was the only structure to obtain convergence. Results were as follows: $F_{(1,24864)} = 257.20, p < .001$ for hypernymy, $F_{(1,11145)} = 110.81, p < .001$ for English fluency, $F_{(1,13823)} = 42.61, p < .001$ for sex, and $F_{(1,24864)} = 2.78, p = .096$ for the interaction between hypernymy and sex.

For concreteness, a heterogeneous autoregressive-1 covariance structure was specified. Results were as follows: $F_{(1,7811)} = 660.68, p < .001$ for concreteness, $F_{(1,11280)} = 107.37, p < .001$ for English fluency, $F_{(1,10084)} = 10.84, p = .001$ for sex, and $F_{(1,7811)} = 4.82, p = .028$ for the interaction between sex and concreteness. For familiarity, a heterogeneous autoregressive-1 covariance structure was specified. Results were as follows: $F_{(1,11610)} = 201.27, p < .001$ for familiarity, $F_{(1,11148)} = 112.77, p < .001$ for English fluency, $F_{(1,11470)} = 81.73, p < .001$ for sex, and $F_{(1,11610)} = 76.89, p < .001$ for the interaction between sex and familiarity.

4.9.4 5-Predictor Models

Because none of the interaction effects were statistically significant in both English fluency and sex interaction models for the Coh-Metrix features, no 5-predictor models were tested.

4.9.5 Summary

In the one predictor models, WordNet polysemy, frequency, familiarity, meaningfulness, and hypernymy of verbs were associated with higher endorsement of depressive symptomology, while Framenet polysemy, concreteness, imageability, Coh-Metrix polysemy, and hypernymy of nouns were associated with lower endorsement of depressive symptomlogy. In particular, hypernymy of verbs appeared to have a large effect relative to the other predictors ($B = .059$), followed by frequency ($B = .027$), and FrameNet polysemy ($B = -.024$). For hypernymy of verbs, the range observed in our data set was .73-4.17, meaning the effect ranged from .04-.25. Thus, having 4 items with specific verbs (or verbs that had approximately 4 superordinate levels of meaning), the composite CES-D score would change by 1 point. When considering the average score for familiarity ($\bar{x} = 591.3$), the effect of familiarity is 1.18, and for a one SD increase, the effect is 1.21. The difference in the effect of familiarity between the minimum and maximum for familiarity score is .1. Items with more familiar words had higher endorsement of depressive symptomology. Similarly, across the range of imageability, there is a .18 difference, such that the more imageable items had responses that were .18 points lower than less imageable items. While these are fairly small effects, they could be meaningful at a composite scale level.

In the three predictor models, the same pattern for direction was present. The coefficient estimates were similar or slightly lower, except for frequency, where coefficient estimate nearly doubled. All features except for meaningfulness and hypernymy for nouns remained statistically significant when controlling for sex and English fluency.

There were statistically significant interaction effects between English fluency and the features of concreteness and hypernymy of nouns. For concreteness, among individuals who were non-English dominant, the predicted CES-D score for the most concrete item was .90 lower than the predicted CES-D score for the least concrete item. For individuals who were very fluent in English, the most concrete item was .45 points lower than the least concrete item. Thus, this suggests that concreteness had a stronger effect for individuals who woere less fluent in English. For hypernymy of nouns, the effect of hypernymy was stronger for individuals who were less fluent in English. For an individual who was not dominant in English, the predicted CES-D score for a item with nouns that had high scores for hypernymy was .26 points lower than for an item that had nouns with low scores for hypernymy. The difference was only .14 points for individuals who were very fluent in English.

There were statistically significant interaction effects between sex and polysemy (WordNet), polysemy (FrameNet), frequency and familiarity. For polysemy (WordNet), the interaction effect was such that the difference between predicted values for men and women (holding all else constant) was less when the words in an item had many possible meanings. In other words, the interaction effect counteracted the negative effect of being male. For

polysemy (FrameNet), the effect of being a male was positive, which was not the case in the 3-predictor model (or in any other models where sex was included). The interaction effect here is accounting for the negative effect that was typically associated with being male, as the main effect for sex is no longer statistically significantly different from zero.

For frequency, the difference in predicted CES-D scores for males and females (holding all else constant) was larger when the frequency score was lower. As the frequency score associated with the words in the item increased, males responses increased. The main effect for the feature was no longer statistically significantly different from zero, suggesting that the effect of word frequency depended on the sex of the individual. The interaction effect for familiarity suggested that male and female responses became more similar as familiarity increased.

4.10 Multiple Feature Models

In order to determine whether these features were *uniquely* contributing to the CES-D responses, we specified a multiple feature model. We selected features that had large effects relative to the other predictors in the domain in the single predictor models based on the coefficient estimates and standard errors. These features were:

1. Word length (Domain: Length)
2. Proportion of Nouns (Domain: Morphology I)
3. Count of bound morphemes (Domain: Morphology II)
4. Proportion of content words (Domain: Morphology II)
5. Relative clauses (Domain: Syntax)
6. Cultural loading (Domain: Semantics I)
7. Positive affect (Domain: Semantics II)
8. Concreteness (Coh-Metrix) (Domain: Semantics II)
9. Hypernymy of nouns (Domain: Semantics II)

4.10.1 Full Feature Model

The full feature model included all of the features. A heterogeneous compound symmetric covariance structure was specified. Results of the tests of fixed effects were as follows: for word length $F_{(1,15497)} = 166.09, p < .001$, for count of nouns $F_{(1,12511)} = 70.10, p < .001$, for bound morphemes $F_{(1,13221)} = 533.88, p < .001$, for proportion of content words $F_{(1,14771)} = .238, p = .626$, for relative clauses $F_{(1,15222)} = 1505.14, p < .001$, for cultural

loading $F_{(1,22551)} = 870.00, p < .001$, for positive affect subscale $F_{(1,19404)} = 196.71, p < .001$, for concreteness $F_{(1,7301)} = .182, p = .669$, and for hypernymy of nouns $F_{(1,11777)} = 12.84, p < .001$.

Content and concreteness were not statistically significant in this model, thus, these predictors were removed from the model and a revised full model was tested.

4.10.2 Full Feature Model – Revised

For the full model with content and concreteness removed, results of the tests of fixed effects were as follows: for word length $F_{(1,17154)} = 228.73, p < .001$, for count of nouns $F_{(1,14443)} = 94.19, p < .001$, for bound morphemes $F_{(1,14205)} = 553.56, p < .001$, for relative clauses $F_{(1,7714)} = 2097.5, p < .001$, for cultural loading $F_{(1,22976)} = 884.59, p < .001$, for positive affect subscale $F_{(1,18061)} = 213.63, p < .001$, and for hypernymy of nouns $F_{(1,12152)} = 18.95, p < .001$.

The coefficient estimates were as follows: holding all other predictors constant, a 1 unit increase in average word length was associated with a .101 increase in CES-D response, a .1 increase in proportion of nouns was associated with a .077 decrease in CES-D response, an increase in the number of bound morphemes was associated with a .165 decrease in CES-D response, the presence of a relative clauses resulted in responses that were .56 higher compared to items without relative clauses, a culturally loaded item was .237 lower on CES-D response than an item that was not culturally loaded, items that were on the positive affect scale were .155 points higher on average than items not on the positive affect subscale, and for a one unit increase in hypernymy of nouns, the average CES-D response decreased by .013. These results are included in table D.34.

4.10.3 Full Feature Model – Revised, with Sex and English Fluency

The final model tested included all of the features included in the revised full model, as well as sex and English fluency. Coefficient estimates are provided in table D.35. Results of the fixed effects were as follows: for word length $F_{(1,17053)} = 230.58, p < .001$, for count of nouns $F_{(1,14343)} = 95.80, p < .001$, for bound morphemes $F_{(1,14157)} = 555.70, p < .001$, for relative clauses $F_{(1,7714)} = 2086.93, p < .001$, for cultural loading $F_{(1,22828)} = 881.91, p < .001$, for positive affect subscale $F_{(1,18078)} = 216.26, p < .001$, for hypernymy of nouns $F_{(1,12141)} = 19.76, p < .001$, for sex $F_{(1,2602)} = 33.12, p < .001$, and for English fluency $F_{(1,2602)} = 26.21, p < .001$.

Results for these models indicate that when controlling for other features as well as sex and English fluency, the features are still statistically significant. In most cases, the coefficient estimates were higher in this full feature model than in the single predictor models.

4.11 Generalized Linear Mixed Models

4.11.1 Feature Selection

As described in the methods section, we selected a set number of features using results from the linear mixed model analyses to essentially "pilot" the generalized linear mixed models. The features chosen and the rationale are described below.

1. Average Word length, letters: This feature had a relatively large effect in each of the models it was included in. For example, in the three predictor model, the effect showed that for each additional increase in the average word length, the CES-D response decreased by .12. Considering that the range for this feature was 4 units, (2.67-6.67) the difference in an item that had longer words on average and one with shorter words on average was -.48, approximately half a point.

2. Morphology-POS: For the morphology features that represented the parts of speech, prepositions tended to have the largest effect in several models considered. Specifically, on average, they increased the CES-D response. In other words, a greater proportion of prepositions was associated with higher reported levels of depressive symptomology. This effect was relatively small when compared to other effects, and will represent a moderate or a small effect.

3. Morphology-Other features: For the morphological features that were separate from parts of speech, the count of bound morphemes had the largest effects. Specifically, an item that had the maximum number of bound morphemes (4) had responses that were -.27 lower than an item that had only one bound morpheme.

4. Relative clauses: the presence of a relative clause showed a fairly large effect in several models considered. In a single predictor model, the coefficient was .42, indicating that for each additional relative clause, the CES-D score increased by .42 (indicating higher endorsement of depressive symptomology).

5. Cultural loading: From the first subset of semantic features, cultural loading appeared to have a large effect on responses. Specifically, in the single predictor model, a culturally loaded item had responses that were -.27 points lower on average than a non-culturally loaded item. There was also a small, but statistically significant interaction effect between English fluency and cultural loading, in the opposite direction.

4.11.2 Model Specification

In determining the type of model to use for the generalized linear mixed models, Kolmogorov-Smirnov tests were obtained to determine if the distribution for each item appeared to be distributed as a Poisson distribution. The results of these preliminary analyses, which were

presented in a poster, indicated that 13 of the 20 items were approximately Poisson distributed. Therefore, a Poisson distribution was specified in the analyses. A logistic link function was also specified. Since compound symmetric covariance structures were used in the majority of cases, this was the type of covariance structure specified for all of the generalized linear mixed models.

4.11.3 Model Fit

In comparing the generalized linear mixed models to the linear mixed models, we considered the AIC and BIC. These allow us to compare models that are not nested. However, we also looked at the coefficient estimates provided for the models which had the same predictors. The major question we sought to answer was whether generalized linear mixed models, and more specifically, a Poisson distribution with a log link-function offered any clear advantages over a linear mixed model.

4.11.4 Comparison of Models

In all cases, the generalized linear mixed models fit our data poorly in comparison to the linear mixed models. There were differences in the AIC and BIC in the order of 20,000. Using a chi-square difference test on one degree of freedom, a value this large is incredibly unlikely. Because we found that the fit was not improved and was in fact made worse, we did not obtain generalized linear mixed models for our entire set of features. The direction of the results from the GLMMs were the same as the LMMs. This is evident when looking at the exponentiated coefficients from the models, available in D.29-D.33. Exponentiated coefficients represent the change in the count when increasing a continuous variable by 1 unit, or when changing the category of a dummy coded variable. Since this is a factor of change, coefficients closer to 1 represent little or no change. Coefficients lower than 1 represent a negative effect, and coefficients greater than 1 represent a positive effect. However, because the fit was poor for these models, while the interpretation is made more complex, we did not see a clear advantage to using them. When comparing predicted estimates for CES-D responses using different levels of the predictors, the one advantage of these models would be that they are bounded at zero. In the linear mixed models, negative predicted values for the CES-D response could occur.

Chapter 5

Discussion

The purpose of the present study was, broadly, to investigate a methodological and theoretical question about the language used on questionnaires. Our methodological question sought to investigate the various ways in which we could describe items on a questionnaire. These descriptions can be thought of as the features of the items, or as some have put it, itemmetrics. In addition, we sought to see whether features that appeared to affect responses continued to have an effect in the presence of other person-level characteristics, specifically sex and English fluency. One way to consider the results of this study is to separate the substantive findings from the more methodological findings. Because we used a single questionnaire, we are not able to make broad claims about these features and how they would impact the level of responses on any given questionnaire. However, our findings about this questionnaire might be of interest to a substantive researcher who uses a depressive symptomology instrument. The methodological aspects of the present study might compel researchers to investigate commonly used questionnaires, or at the very least to consider the myriad ways in which language can be described.

5.1 Summary of Results: Linguistic Features of the CES-D

In this section, we will summarize the results of our linguistic analysis as applied to the CES-D. Using previous work in itemmetrics as well as linguistic theory, we proposed that linguistic properties could be organized under three domains: morphological, syntactic, and semantic. We then explored the numerous ways in which a single property could be coded into various features and applied these features to the 20 items on the CES-D. Through this exploration, we found that linguistic theory is a viable framework in which to organize linguistic properties. We did find that some features did not vary across CES-D items, or that some correlated so highly that it would be redundant to include both in the further models we obtained. If we had used a different set of items for the testing of these linguistic features, the results may have varied. This linguistic analysis is a major

contribution of the present study since up until this point, there has not been a study which has sought to describe the CES-D items in terms of such a wide range of features. In fact, to our knowledge, there has not been a study that has proposed a set of features organized into the domains described here and then applied to a self-report measure of depressive symptomology.

5.2 Summary of Results: Empirical Analysis of Responses to CES-D Items

In this section, we will summarize the results in terms of the type of model and the direction of the effect on CES-D responses. By type of model, we are referring to the number of predictors in the model. A single predictor model examines the feature alone, while the 3-predictor models examined the effect of the feature when controlling for sex and English fluency. The interaction models examine how the feature interacts with these characteristics.

Initially, we started with a list of 80 different features and through the correlational analysis, we trimmed our feature set down to 37 distinct features.

5.2.1 1-Predictor Models

Within the 1-predictor models, 13 of the total 37 features were found to have positive coefficients which were statistically significant. These features were (by domain):

- (Domain: Length) word count
- (Domain: Morphology I) proportion of main verbs, proportion of adverbs, proportion of pronouns, proportion of prepositions
- (Domain: Syntax) relative clauses
- (Domain: Semantics I) positive affect subscale, somatic subscale
- (Domain: Semantics II) WordNet polysemy, word frequency (all words), familiarity, meaningfulness, and hypernymy of verbs

Negative effects: Within the 1-predictor models, 23 of the total 37 features were found to have negative coefficients which were statistically significant. These features were (by domain):

- (Domain: Length) average word length in letters
- (Domain: Morphology I) proportion of nouns, proportion of adjectives, proportion of copular verbs
- (Domain: Morphology II) bound morphemes, suffixes, proportion of content words

- (Domain: Syntax) ambiguity, negations, possessives,
- (Domain: Semantics I) specificity, concreteness (as manually rated), cultural loading, gender loading, depressed affect subscale, interpersonal problems subscale, referring to others
- (Domain: Semantics II) FrameNet polysemy, concreteness (Coh-Metrix), imageability, Coh-Metrix polysemy, and hypernymy of nouns

5.2.2 3-Predictor Models

Of the 13 features that had statistically significant positive effects in the single predictor model, 11 remained statistically significant when controlling for sex and English fluency. However, English fluency was statistically significant in only 7 of the 11 models. These features included (by domain):

- (Domain: Length) number of words
- (Domain: Morphology I) proportion of main verbs, proportion of pronouns*, prepositions*, relative clauses*
- (Domain: Semantics I) positive affect, somatic subscale*
- (Domain Semantics II) WordNet polysemy, frequency, familiarity, and hypernymy of verbs*

Of the 23 features that had statistically significant negative effects in the single predictor model, 17 continued to be statistically significant when controlling for sex and English fluency. Within those 17, however, there were only 9 models in which sex and English fluency were statistically significant. These features included (by domain):

- (Domain: Length) average word length
- (Domain: Morphology I) nouns, proportion of adjectives*, copular verbs
- (Domain: Morphology II) bound morphemes, suffixes*, proportion of content words*
- (Domain: Syntax) ambiguity, negations*, possessives*
- (Domain: Semantics I) specificity*, concreteness*, culture, gender*
- (Domain: Semantics II) FrameNet polysemy, concreteness, imageability*, and hypernymy for nouns

In sum, in the 3-predictor models, 28 of the original 37 features continued to have statistically significant effects. Of these, there were 16 feature models where sex and English fluency were also statistically significant. Interestingly, sex was a statistically significant predictor in all models, whereas English fluency was not.

** indicates that these models did not have statistically significant effects for English fluency*

5.2.3 4-Predictor English fluency by Feature Interaction Models

The models that had statistically significant effects for the feature, English fluency, and sex, were tested to determine whether there was an interaction between the feature and English fluency. In some cases, models that did not have statistically significant effects for English fluency were included in these interaction models as well.

There were 8 features which were included for these analyses and had positive effects for the main effect of the feature. These features were (by domain):

- (Domain: Length) number of words*
- (Domain: Morphology I) proportion of main verbs*, proportion of prepositions*
- (Domain: Syntax) relative clause
- (Domain: Semantics I) positive affect subscale
- (Domain: Semantics II) WordNet polysemy*, frequency*, and familiarity*

There were 13 features that continued to show negative effects when an interaction term was added. These features included (by domain):

- (Domain: Length) average word length in letters
- (Domain: Morphology I) adjectives*, copular verbs*
- (Domain: Morphology II) bound morphemes
- (Domain: Syntax) ambiguity*, negations*, possessives*
- (Domain: Semantics I) culture, interpersonal problems subscale
- (Domain: Semantics II) FrameNet polysemy*, concreteness, and hypernymy for nouns

** indicates that these models did not have statistically significant effects for the interaction term* Of the 21 features tested, 8 features had statistically significant interaction terms.

5.2.4 4-Predictor Sex by Feature Interaction Models

In the 4 predictor models, we observed that the same features continued to have positive effects. Five of these features had statistically significant interaction terms. The features that had statistically significant effects included (by domain):

- (Domain: Length) number of words*
- (Domain: Morphology I) main verbs, prepositions*
- (Domain: Morphology II) relative clause*
- (Domain: Semantics I) positive affect
- (Domain: Semantics II) WordNet polysemy, frequency, and familiarity

In addition, the same features continued to have negative effects. These features included (by domain):

- (Domain: Length) word length in letters
- (Domain: Morphology I) nouns*, adjectives, copular verbs
- (Domain: Morphology II) bound morphemes
- (Domain: Syntax) ambiguity, possessives
- (Domain: Semantics I) culture*, interpersonal problems
- (Domain: Semantics II) FrameNet polysemy, concreteness, and hypernymy for nouns

Ten of these features had statistically significant interaction terms ** indicates that these models did not have statistically significant effects for the interaction term*

5.2.5 5-Predictor Models

Models that had statistically significant interaction terms for sex by feature as well as sex by English fluency. For positive effects, there were no 5-predictor models that were fit.

For negative effects, average word length in letters was included, and the sex by feature interaction term was just narrowly non-significant. Bound morphemes (Domain: Morphology II), possessives (Domain: Syntax) and interpersonal problems (Domain: Semantics I) were included and had statistically significant interaction terms.

5.2.6 Multiple Predictor Model

A small number of multiple predictor models were obtained which included the features of word length, nouns, bound morphemes, relative clauses, cultural loading, positive affect, and hypernymy of nouns. In these models, we found that all features were statistically significant. In an additional model which controlled for sex and English fluency, all features continued to be statistically significant, in addition to sex and English fluency. Therefore, this is a promising direction for future research.

5.2.7 Generalized Linear Mixed Model

A sensitivity analysis was conducted to determine whether assuming that the outcome variable was drawn from a non-normal distribution would be advantageous for the results. In comparing the model fit, the generalized linear mixed model which specified a Poisson distribution had a less optimal fit. In addition, the results of the statistical tests and their corresponding interpretations did not differ meaningfully between the generalized linear mixed effects models and the linear mixed effects models. Therefore, given the increased complexity in interpreting generalized linear mixed effects models, we chose to use linear mixed effects models for our modelling and interpretation of the effects of linguistic features and individual characteristics on item responses.

5.3 Linkage to Research Questions

We sought to explore several research questions in this study. Here, we discuss how our findings relate to each research question.

5.3.1 What are the different ways to operationalize linguistic properties within the framework of psychometrics? How are these operationalizations correlated?

Through our enumeration of linguistic features, we demonstrated that there are many potential itemmetrics for questionnaire items, 80, to be precise. Significantly, drawing from the work done by Goldberg (1963, 1968); Payne (1974), we were able to generate itemmetrics for our instrument that were completely independent of the data collected. In other words, we have attempted to expand on the approaches used by Goldberg (1968), in which the lexicographic indices were the only itemmetrics that were not dependent on having one or multiple samples of data. In addition, we also used ratings from judges to characterize the items, which was one methodological approach seen in previous work (Payne, 1974).

Moreover, our use of open-source language processing tools for this purpose appears to be novel. Our approach is rooted in linguistic theory and uses tools developed by computer scientists. This, we believe, is a large improvement over the itemmetrics originally used in

the literature, which depended largely on dictionaries. On one hand, the development of these features and explanation of how the tools can be used serves as a "proof of concept" for other researchers who are interested in this area of research.

Our study addresses the correlations among various linguistic properties as well, and this also informed our decisions about which features to include within our analyses. For example, we found that simply counting the various types of parts of speech within an item would lead to strong correlations among the various parts of speech features. Thus, we opted to use the proportions of the parts of speech, which are able, to an extent, to remove the confounding effect of the number of words in an item. Furthermore, our comparison of manually coded semantic features and features derived from Coh-Metrix showed that there was variability in how an automated approach might assign values for certain features (e.g., concreteness) compared to individuals rating the items. In making suggestions for future research in this area, we think that a sound approach would be to use both methods, depending on the resources available to the researcher. For example, a researcher who does not have the resources to have items rated by trained judges, an automated approach could still offer some insight about how certain features affect responding. We would recommend that a researcher who does have the resources to have items rated by trained judges to also make use of the automated tools available and to of course compare the results from either approach.

The takeaway for this research question is that technology has removed potential barriers that have likely kept itemmetric research from expanding. We urge other researchers to make use of these tools in order for our understanding of itemmetrics to grow. Applying these itemmetrics to response options is a potential next step for this program of research, as well as using itemmetrics for other instruments.

5.3.2 Are linguistic properties/features related to higher/lower endorsement of depressive symptomatology? Which ones?

This research question is more substantive in nature and was addressed using 1-predictor models. Our findings reveal that item features do indeed lead to higher and lower endorsement of depressive symptomology. Generally, there were more features that led to lower endorsement of depressive symptomology. To link these findings to earlier theoretical work about response sets such as Paulhus (1991), Bentler et al. (1971), and Kulas and Stachowski (2013), our findings could potentially be interpreted as challenging the standard conceptualization of response sets. That is, response sets are typically considered to be a characteristic of the individual responding to the item. Instead, we argue that our evidence suggests that properties of items themselves could affect the way that all individuals respond, on average. The present study did not address response styles specifically, but we thought that certain findings could contribute to that line of research. For example, our finding that culturally loaded items lead to lower endorsement of depressive symptomology suggests that there

is perhaps a social desirability bias inherent in certain items, leading individuals to not endorse such items.

On the other hand, our findings introduce several new questions for exploration. For example, we found that items which had higher scores for concreteness (both types) and hypernymy of nouns (i.e., specificity of meaning) led to lower responses on average. Conversely, we found that polysemy, which can be thought of as a characterization of ambiguity, also led to lower endorsement of depressive symptomology. It is unclear why features that seem to suggest greater specificity or tangibility of the item have similar effects as features that suggest ambiguity of meaning in the item. This seeming contradiction points to the importance of considering the effects of multiple features in a single model.

5.3.3 Which properties are most related to higher/lower endorsement of depressive symptomatology?

Generally, we considered word length, nouns, bound morphemes, proportion of content words, relative clauses, cultural loading, positive affect, and hypernymy to have the strongest effects. Relative clauses and positive affect had positive effects on responses while the other features had negative effects.

5.3.4 Does language background and/or sex interact with certain item properties to lead to higher/lower endorsement of depressive symptomatology?

Using 3-predictor models allowed us to determine whether features had statistically significant effects on CES-D responses when controlling for English fluency and sex. These models suggested that the features, in several cases, were more predictive of the responses than English fluency. This was not the case for sex, however. Thus, some features might better explain the CES-D responses than English fluency.

On the other hand, we did find that several features had statistically significant interactions with English fluency, though there did not appear to be a pattern as to which domains of features interacted most with English fluency. Some of the features that interacted significantly with English fluency were more descriptive or related to the mechanics of language/language processing, (e.g., average word length, bound morphemes), while others were related to semantics (e.g., cultural loading, concreteness, positive affect). We think these results suggest that the construct of English fluency has several components, some related to the mechanics of using English, and others related to interpreting English, especially within a particular cultural context.

On the other hand, the results suggested that sex was able to uniquely predict CES-D responses, as the 3 predictor models showed that including sex did not result in non-

significance for the feature or sex. Moreover, several of the features had statistically significant interactions with sex.

Generally, we think that there is evidence that certain linguistic features interact with individual characteristics. Our use of linear mixed effects models revealed that both individual characteristics and features of the items, and their interactions, are predictive of CES-D responses. In terms of the size of the effects, they are not large enough to threaten the utility of the instrument we used, but we think that researchers should be aware that individual characteristics can affect responding in potentially meaningful ways. The take-away message for researchers is that scales do not simply measure the construct of interest. Here we have modelled responses as being predicted by both features of the items and individual characteristics, and their interactions, and therefore, this "measurement error" does not appear to be random. It appears to be attributable to itemmetrics and individual characteristics.

5.4 General Conclusions and Interpretations

In responding to the question of whether linguistic features impact responses to a problematic extent, the answer is not completely clear. Most of the features had modest effects. However, these results beg the question of whether these small effects are so minute as to be excusable. For a substantive researcher, it might be the case that considering these small effects is too burdensome to take on when attempting to investigate questions about treatment or development. For a more methodologically-focused researcher, particularly one who is interested in cognitive psychology, however, these results may be more intriguing. This is because it is challenging to explain the reason why some features lead to higher responses, such as relative clauses, while others lead to lower responses, such as average word length. At the lower levels of linguistic analysis, it is not completely clear why a certain feature would influence responses in a certain direction. At the same time, it is worth asking the question of why one would expect the features to have no effect on responses.

The results of this investigation, therefore, might be useful in creating new directions for a methodological researcher who is interested in measurement or the language used on questionnaires. A crucial next step would be to examine multiple features simultaneously and perhaps tie certain features into a processing framework. Within a processing framework, it would be fascinating to explore possible explanations for why combinations of features might increase the processing load and therefore lead to higher or lower endorsement of depressive symptomology, or whichever construct is being measured. This is a quite challenging question. There may be an argument for why certain features would increase the processing load, such as longer words or more function words like prepositions. However, connecting that higher cognitive demand to a particular level of response is murky. Perhaps a lower response is chosen because it is physically closer to the item stem or requires less

reading of the options. Or, in the opposite direction, perhaps the high processing load leads to the individual overestimating their response as a way to cut down on additional fatigue. If one were to examine this within the scope of this questionnaire, it might make sense to look at the reverse-coded items vs. the non-reverse coded items. Treating the responses to the reverse-coded items as they originally appeared on the questionnaire (that is, a response of 0 is considered a 0, rather than a 3 as it would once the responses were reverse-coded), one might consider whether a higher processing load leads to lower responses as a function of physical distance from the item stem.

An alternative way to consider this question would be to look at a questionnaire measuring a different construct, but that has a similar response scale. If the patterns were similar, there would be a strong case for combinations of linguistic features contributing construct-irrelevant variance. However, that situation is quite a simplistic view of processing. Considering item stems that deal with the topic of positive or negative affect might have a higher cognitive demand at baseline when compared to a personality questionnaire.

5.5 Limitations and Future Directions

There are a number of limitations to this study, some of which could be addressed in additional studies, and others that are more difficult to remedy. Broadly, as an exploratory study, the results here are in no way definitive. There is evidence that linguistic features affect responses, but more research is required for any generalizable conclusions to be drawn.

As an exploratory study, there are caveats to consider when interpreting our findings. We examined a single measure, the CES-D, which means our findings at this point are in the context of that measure alone. Applying the linguistic analysis and response modelling we propose here to other measures would strengthen the inferences we can draw from our results. Only after replication studies have been conducted can we begin to discuss these findings as being applicable to a wider range of instruments.

With regard to our linguistic analysis, although we covered multiple domains (morphology, syntax, and semantics), we only examined a limited set of features. Other properties and features could have been included in our initial set of 80, but some were not chosen for analysis. Our rationale for choosing certain features over others is that there were some features (for example, some of those provided by Coh-Metrix) which did not show any variability across the 20 items. If a different instrument was used, it would be likely that some of the features shown here would not vary across items. Moreover, it could be the case that new features would need to be explored. We did, however, ensure that our features covered a range of linguistic domains.

A complex issue is that we have limited our analysis to the linguistic rules and usage of English. Some of these features do not *exist* in other languages. Therefore, even when considering the CES-D, which has been translated into a number of languages, our approach

here would have to be modified in terms of the language of interest. More substantively, it is difficult to say whether our findings, particularly with the semantic features, are able to be interpreted as being related to language alone. In particular, separating the construct of depressive symptomology from attributes of the language used to describe depressive symptom was not addressed here.

In terms of our methodological approach, a limitation is the unitary nature of the linguistic analysis, especially in the case of structural features. Efforts were made to include semantic features as well and to examine relationships across the features used. However, it was still challenging to integrate across the many features. We argue that we were successful in our purpose to first enumerate many features, then to identify candidate features for inclusion in our modelling of CES-D item responses. Through this work, however, we've seen that defining many features can make interpretation more challenging.

In terms of modelling the role of item features, both alone and in combination with individual characteristics, on people's responses to items, there are a number of limitations. The fact that most of our models only considered one feature at a time could be considered as one limitation. However, we did examine a few multiple-feature models, which suggested that the various features were contributing uniquely to the CES-D responses. Future work can further examine the simultaneous role of multiple linguistic features in a single model. These groupings of linguistic features could then also be examined with individual characteristics. One of our focal individual characteristics in this study was English fluency; the manner in which English fluency was modelled may be considered to be a weakness by some. Specifically, the 4-level measure used for English fluency was analyzed as a quantitative predictor. Future studies could consider treating this variable as a categorical predictor, or could use a different measure for English fluency which more closely approximates an interval scale level of measurement. Additionally, a measure that provides more potential for variability in English fluency may strengthen the results. More generally, we only considered two individual characteristics, sex and English fluency, in our models. Including other individual characteristics, such as cultural background, personality traits, or age, could be more informative. To do this, however, we would need a sample that had a wider age range. To some extent, our sample, which consisted of university students, constrained our ability to include other individual characteristics in the study. A sample that was more representative of the general adult population would be useful. A clinical sample could perhaps provide additional insight.

Despite these shortcomings, the present study makes a novel methodological contribution. That is, we have defined a number linguistic features which range from those which are structural (e.g., morphological and syntactic features) to those which are semantic. In addition, we have empirically demonstrated that some features of the items on the CES-D may affect how individuals respond. They also may interact with characteristics of individuals. This exciting work suggests many interesting future directions and points to a situation in

which an interdisciplinary approach for improving our instruments is warranted. This limitation could be more fully addressed if multiple researchers were working on this same topic, especially if an interdisciplinary approach was taken via a collaboration among psychologists, cognitive scientists, and linguists. The generation of more interest would contribute significantly to psychometrics by improving our instruments and better understanding how individuals respond to questionnaires. We see this project as a unique contribution toward that end. From a practical standpoint, a barrier to generating interest in this area is that many researchers might not be primarily concerned with how the instruments they use work. The lack of a cohesive research program in itemmetrics is perhaps reflective of the fact that researchers are more interested in the production of other types of knowledge. A future challenge, therefore, is gaining buy-in from the many psychologists who rely on self-report questionnaires. If psychologists are not concerned with this area of research and do not use the results to inform their own research, itemmetrics could remain buried in the literature. This is not an issue of whether this topic is important, but whether it is perceived as important to researchers who could benefit most from new knowledge in this area. More optimistically, there is a unique opportunity for other quantitatively inclined researchers to focus on itemmetrics.

References

Ace, M. C., & Davis, R. V. (1973). Item structure as a determinant of item difficulty in verbal analogies. *Educational and Psychological Measurement*, *33*, 143–149.

Allport, G. W., & Odbert, H. S. (1936). *Trait names: A psycho-lexical study* (Vol. 47). doi: 10.1037/h0093360

Angleitner, A., John, O. P., & Löhr, F.-J. (1986). It's what you ask and how you ask it: An itemmetric analysis of personality questionnaires. In *Personality assessment via questionnaires. current issues in theory and measurement.* (pp. 61–108).

Archer, R. P., & Gordon, R. (1994). *Psychometric stabliity of MMPI-A item modifications* (Vol. 62) (No. 3).

Bailin, A., & Grafstein, A. (2001). The linguistic assumptions underlying readability formulae: a critique. *Language and Communication*, *21*(3), 285–301. doi: 10.1016/S0271-5309(01)00005-2

Bain, R. (1931). Stability in questionnaire response. *American Journal of Sociology*, *37*(3), 445–453.

Baker, C. F., Fillmore, C. J., Lowe, J. B., Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of the 36th annual meeting on association for computational linguistics* - (Vol. 1, p. 86). Morristown, NJ, USA: Association for Computational Linguistics. Retrieved from `http://portal.acm.org/citation.cfm?doid=980845.980860` doi: 10.3115/980845.980860

Barnette, J. J. (2000). Effects of stem and likert response option rerversals on survey internal inconsistency: If you feel the need, there is a better alternative to using those negatively worded items. *Educational and Psychological Measurement*, *60*(3), 361–370.

Baxter, J. C., & Morris, K. L. (1968). Item ambiguity and item discrimination in the MMPI. *Journal of Consulting and Clinical Psychology*, *32*(3), 309–313. doi: 10.1037/h0025890

Beiser, M., & Hou, F. (2001). Language acquisition, unemployment and depressive disorder among Southeast Asian refugees: A 10-year study. *Social Science and Medicine*, *53*(10), 1321–1334. doi: 10.1016/S0277-9536(00)00412-3

Bejar, I. I. (1993). A generative apprach to psychological and educational measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (Vol. 53, pp. 1689–1699). Hillsdale, NJ: Lawrence Erlbaum Associates. doi: 10.1017/CBO9781107415324.004

Benson, J., & Crocker, L. (1979). Effects of item format and reading ability on objective test performance: a question of validity. *Educational and Psychological Measurement*, *39*, 381–387.

Bentler, P. M., Jackson, D. N., & Messick, S. (1971). Identification of content and style: A two-dimensional interpretation of acquiescence. *Psychological Bulletin*, *76*(3), 186–204. doi: 10.1037/h0031474

Benton, A. L. (1935). The interpretation of questionnaire items in a personality schedule. *Archives of Psychology (Columbia University)*(190), 38.

Bernreuter, R. G. (1935). *Manual for the personality inventory.*

Birenbaum, M., Tatsuoka, K. K., & Gutvirtz, Y. (1992). Effects of response format on diagnostic assessment of scholastic achievement. *Applied Psychological Measurement*, *16*(4), 353–363. doi: 10.1177/014662169201600406

Bond, J. A. (1987). The process of responding to personality items: Inconsistent responses to repeated presentation of identical items. *Personality and Individual Differences*, *8*(3), 409–417. doi: 10.1016/0191-8869(87)90042-0

Broen, W. E. (1960). Ambiguity and discriminating power in personality inventories. *Journal of Consulting Psychology*, *24*(2), 174–179. doi: 10.1037/h0046163

Brown, C., Schale, C., & Nilson, J. E. (2010). Vietnamese immigrant and refugee women's mental health: An examination of age of arrival, length of stay, income and English language proficiency. *Journal of Multicultural Counseling and Development*, *38*(2), 66–.

Camilli, G., Briggs, D. C., Sloane, F. C., & Chiu, T.-W. (2013). Psychometric perspectives on test fairness: Shrinkage estimation. *APA handbook of testing and assessment in psychology, Vol. 3: Testing and assessment in school psychology and education.*, *3*, 571–589. Retrieved from `http://ezproxy.library.capella.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true{&}db=pzh{&}AN=2012-22487-027{&}site=ehost-live{&}scope=site` doi: 10.1037/14049-027

Carlson, J. F. (2013). Clinical and counseling testing. *APA handbook of testing and assessment in psychology, Vol. 2: Testing and assessment in clinical and counseling psychology.*, *2*, 3–17. Retrieved from `http://proxy.lib.sfu.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true{&}db=pzh{&}AN=2012-22486-001{&}site=ehost-live` doi: 10.1037/14048-001

Chang, L. (1995). Connotatively consistent and reversed connotatively inconsistent items are not fully equivalent: generalizability study. *Educational and Psychological Measurement*, *55*(6), 991–997.

Chung, R. C., & Kagawa-Singer, M. (1993). Predictors of psychological distress among southeast Asian refugees. *Social Science & Medicine*, *36*(5), 631–639. doi: 10.1016/0277-9536(93)90060-H

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, *33*(4), 497–505. doi: 10.1080/14640748108400805

Condon, L., Ferrando, P. J., & Demestre, J. (2006). A note on some item characteristics related to acquiescent responding. *Personality and Individual Differences*, *40*, 403–407. doi: 10.1016/j.paid.2005.07.019

Congress, U. (2001). *Public Law 107-110: The no child left behind act of 2001.* Retrieved from `http://www2.ed.gov/policy/elsec/leg/esea02/index.html`

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings.* Boston: Houghton Mifflin.

Cronbach, L. J. (1942). Studies of acquiescence as a factor in the true-false test. *The Journal of Educational Psychology*, *10*(6), 401–415. doi: 10.1037/h0054677

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, *6*(4), 475–494. Retrieved from `http://epm.sagepub.com/content/6/4/475{%}5Cnhttp://epm.sagepub.com/content/6/4/475.full.pdf{%}5Cnhttp://epm.sagepub.com/content/6/4/475.refs` doi: 10.1177/001316444600600405

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models : a generalized linear and nonlinear approach.* Springer.

Dudycha, A. L., & Carpenter, J. B. (1973). Effects of item format on item discrimination and difficulty. *Journal of Applied Psychology*, *58*(1), 116–121. Retrieved from `http://content.apa.org/journals/apl/58/1/116` doi: 10.1037/h0035197

Ebel, R. L. (1956). Obtaining and reporting evidence on content validity. *Educational and Psychological Measurement*, *16*, 268–282.

Ebesutani, C., Drescher, C. F., Reise, S. P., Heiden, L., Hight, T., Damon, J. D., & Young, J. (2012). The Loneliness Questionnaire âĂŞ Short Version : An evaluation of reverse-worded and non-reverse-worded items via item response theory. *Journal of Personality Assessment*, *94*(4), 427–437. doi: 10.1080/00223891.2012.662188

Edwards, A. L. (1953). The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *Journal of Applied Psychology*, *37*(2), 90–93. Retrieved from `http://web.stcloudstate.edu/jtkulas/Edwards1953.pdf{%}5Cnhttp://content.apa.org/journals/apl/37/2/90` doi: 10.1037/h0058073

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research.* Ft Worth, 1957: Dryden Press.

Eignor, D. R. (2013). The standards for educational and psychological testing. *APA handbook of testing and assessment in psychology: Test theory and testing and assessment in industrial and organizational psychology*, *1*, 245–250. Retrieved from `http://psycnet.apa.org/books/14047/013` doi: 10.1037/14047-013

Eisenberg, P. (1941). Individual interpretation of psychoneurotic inventory items. *The Journal of General Psychology*, *25*, 19–40.

Elosua, P., & Lopez-Jauregui, A. (2007). Potential sources of differential item functioning in the adaptation of tests. *International Journal of Testing*, *7*(1), 39–52. doi: 10.1207/s15327574ijt0701_3

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*(3), 380–396. doi: 10.1037/1082-989X.3.3.380

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, *64*(4), 407–433. doi: 10.1007/BF02294564

Enright, M. K., Morley, M., & Sheehan, K. M. (2002). Items by design : the impact of systematic feature variation on item statistical characteristics. *Applied Measurement in Education*, *15*(1), 49–74.

Esposito, J. L., & Jobe, J. B. (1991). A general model of the survey interaction process. *Bureau of the Census Seventh Annual Research Conference Proceedings*, 537–567. Retrieved from `http://www.bls.gov/osmr/pdf/st910040.pdf`

Flesch, R. (1948). A New Readability Yardstick. *The Journal of Applied Psychology*, *32*(3), 221–233. doi: 10.1037/h0057532

Fontenelle, T. (2012). WordNet, FrameNet and other semantic networks in the international journal of lexicography - The net result? *International Journal of Lexicography*, *25*(4), 437–449. doi: 10.1093/ijl/ecs027

Frank, B. (1935). Stability of questionnaire response. *The Journal of Abnormal and Social Psychology*, *30*(3), 320–324.

Glaser, R. I. U. (1949). A methodological analysis of the inconsistency of response to test items. *Educational and Psychological Measurement*, *9*(4), 727–739.

Goldberg, L. R. (1963). A model of item ambiguity in personality assessment. *Education*, *XXIII*(3).

Goldberg, L. R. (1968). The interrelationships among item characteristics in an adjective check list: The convergence of different indices of item ambiguity. *Educational and Psychological Measurement*, 273–296.

Goldberg, L. R., & Kilkowski, J. M. (1985). The prediction of semantic consistency in self-descriptions: characteristics of persons and of terms that affect the consistency of responses to synonym and antonym pairs. *Journal of personality and social psychology*, *48*(1), 82–98. doi: 10.1037/0022-3514.48.1.82

Gorin, J. S. (2006). Item diffficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, *30*(5), 394–411. Retrieved from `http://apm.sagepub.com/cgi/doi/10.1177/0146621606288554` doi: 10.1177/0146621606288554

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202. doi: 10.3758/BF03195564

Greenberger, E., Chen, C., Dmitrieva, J., & Farruggia, S. P. (2003). Item-wording and the dimensionality of the Rosenberg Self-Esteem Scale: Do they matter? *Personality and Individual Differences*, *35*(6), 1241–1254. doi: 10.1016/S0191-8869(02)00331-8

Gronlund, N. E., & Linn, R. L. (1990). *Measurement and evaluation in teaching* (6th ed.). New York: Macmillan.

Haladyna, T. M., & Shindoll, R. R. (1989). Item shells: A method for writing effective multiple-choice test items. *Evaluation & the Health Professions*, *12*(1), 97–106. doi: 10.1177/1056492611432802

Hamby, T., & Ickes, W. (2015). Do the readability and average item length of personality scales affect their reliability? Some meta-analytic answers. *Journal of Individual Differences*, *36*(1), 54–63. doi: 10.1027/1614-0001/a000154

Harasym, P. H., Price, P. G., Brant, R., Violato, C., & Lorscheider, F. (1992). Evaluation of negation in stems of multiple-choice items. *Evaluation & the Health Professions*, *15*(2), 198–220. doi: 0803973233

Hendrickson, G. (1934). Some assumptions involved in personality measurement. *The Journal of Experimental Education*, *2*(3), 243–249.

Hertzman, M., & Gould, R. (1939). The functional significance of changed responses in a psychoneurotic inventory. *Journal of Abnormal and Social Psychology*, *XIV*, 336–350.

Hewitt, M. A., & Homan, S. P. (2004). Readability level of standardized test items and student performance: The forgotten validity variable. *Reading Research and Instruction*, *43*(2), 1–16. Retrieved from `http://ezproxy.usherbrooke.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true{&}db=psyh{&}AN=2004-21944-001{&}lang=fr{&}site=ehost-live{&}scope=site` doi: 10.1080/19388070409558403

Hinton, L. W., Tiet, Q., Tran, C. G., & Chesney, M. (1997). Predictors of depression among refugees from Vietnam: A longitudinal study of new arrivals. *The Journal of Nervous & Mental Disease*, *185*(1), 39–45.

Holden, R. R., Fekken, G. C., & Jackson, D. N. (1985). Structured personality test item characteristics and validity. *Journal of Research in Personality*, *19*(4), 386–394. doi: 10.1016/0092-6566(85)90007-8

Homan, S., Hewitt, M., & Linder, J. (1994). The development and validation of a formula for measuring single-sentence test item readability. *Journal of Educational Measurement1*, *31*(4), 349–358.

Huang, S.-L., & Spurgeon, A. (2006). The mental health of Chinese immigrants in Birmingham, UK. *Ethnicity & Health*, *11*(4), 365–87. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/17028082` doi: 10.1080/13557850600824161

Hughes, H. H., & Trimble, W. E. (1965). The use of complex alternatives in multiple choice items. *Educational and Psychological Measurement*, *XXV*(1), 117–126. doi: 10.1016/S0953-7562(96)80222-X

Huttenlocher, J. (1962). Some effects of negative instances on the formation of simple concepts. *Psychological Reports*, *11*, 35–42.

Jackson, D. N., & Messick, S. (1958). Content and style in personality assessment. *Psychological bulletin*, *55*(4), 243–252. doi: 10.1037/h0045996

Jensen, S. a., Fabiano, G. a., Lopez-Williams, A., & Chacko, A. (2006). The reading grade level of common measures in child and adolescent clinical psychology. *Psychological assessment*, *18*(3), 346–352. doi: 10.1037/1040-3590.18.3.346

Jobe, J. B. (2010). Models and methods Cognitive psychology and self-reports : Models and Methods. *Cognitive Psychology*, *12*(3), 219–227.

Jones, R. R., & Goldberg, L. R. (1967). Interrelationships among personality scale parameters: item response stability and scale reliability. *Educational and Psychological Measurement*, *27*, 323–333.

Khawaja, N. G. (2007). An investigation of the psychological distress of Muslim migrants in Australia. *Journal of Muslim Mental Health*, *2*(1), 5–20. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/15564900701238468{%}5Cnhttp://www.tandfonline.com/doi/full/10.1080/15564900802487519 doi: 10.1080/15564900701238468

Kim, S. H. O., Ehrich, J., & Ficorilli, L. (2012). Perceptions of settlement well-being, language proficiency, and employment: An investigation of immigrant adult language learners in Australia. *International Journal of Intercultural Relations*, *36*(1), 41–52. Retrieved from http://dx.doi.org/10.1016/j.ijintrel.2010.11.010 doi: 10.1016/j.ijintrel.2010.11.010

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. *Technical Training*, *Research B*(February), 49. Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/a006655.pdf doi: ERIC#:ED108134

Kingston, N. M., Scheuring, S. T., & Kramer, L. B. (2013). Test development strategies. *APA handbook of testing and assessment in psychology, Vol. 1: Test theory and testing and assessment in industrial and organizational psychology.*, *1*, 165–184. Retrieved from http://content.apa.org/books/14047-009 doi: 10.1037/14047-009

Klare, G. R. (1974). Assessing readability. *Reading Research Quarterly*, *10*(1), 62–102.

Kulas, J. T., & Stachowski, A. A. (2012). Social desirability in personality assessment: A variable item contamination perspective. *The International Journal of Educational and Psychological Assessment*, *11*(1), 23–42.

Kulas, J. T., & Stachowski, A. A. (2013). Respondent rationale for neither agreeing nor disagreeing: Person and item contributors to middle category endorsement intent on Likert personality indicators. *Journal of Research in Personality*, *47*(4), 254–262. Retrieved from http://dx.doi.org/10.1016/j.jrp.2013.01.014 doi: 10.1016/j.jrp.2013.01.014

Kuncel, N., & Tellegen, A. (2009). A conceptual and empirical reexamination of the measurement of the social desirability of items. *Personnel Psychology*, *62*(2), 201–228. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1744-6570.2009.01136.x/full doi: 10.1111/j.1744-6570.2009.01136.x

Kuncel, R. B., & Fiske, D. W. (1974). Stability of response process and response. *Educational and Psychological Measurement*, *34*, 743–755.

Lakoff, G. (1987). *Women, fire, and dangerous things : what categories reveal about the mind*. University of Chicago Press.

Lane, S. (1991). Use of restricted item response models for examining item difficulty ordering and slope uniformity. *Journal of Educational Measurement*, *28*(4), 295–309.

Lee, S., Choi, S., & Matejkowski, J. (2013). Comparison of major depressive disorder onset among foreign-born Asian Americans: Chinese, Filipino, and Vietnamese ethnic groups. *Psychiatry Research*, *210*(1), 315–322. Retrieved from http://dx.doi.org/10.1016/j.psychres.2013.03.030 doi: 10.1016/j.psychres.2013.03.030

Lentz, T. F. (1934). Reliability of opinionaire technique studied intensively by the retest method. *Journal of Social Psychology*, *V*, 338–364. doi: 10.1017/CBO9781107415324.004

Lentz, T. F. (1938). Acquiescence as a factor in the measurement of personality. *Psychological Bulletin*, *35*, 659.

Lenzner, T. (2014). Are readability formulas valid tools for assessing survey question difficulty? *Sociological Methods & Research*, *43*(4), 677–698. Retrieved from http://smr.sagepub.com/cgi/doi/10.1177/0049124113513436 doi: 10.1177/0049124113513436

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*(3), 635–694. doi: 10.2466/pr0.1957.3.3.635

McHugh, R. K., & Behar, E. (2009). Readability of self-report measures of depression and anxiety. *Journal of consulting and clinical psychology*, *77*(6), 1100–12. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19968386 doi: 10.1037/a0017124

McHugh, R. K., Rasmussen, J. L., & Otto, M. W. (2011). Comprehension of self-report evidence-based measures of anxiety. *Depression and Anxiety*, *28*(7), 607–614. doi: 10.1002/da.20827

Mckeown, K. R., & Radev, D. R. (1999). Collocations. In H. Dale, Robert, Moisl, Hermann, Somers (Ed.), *A handbook of natural language processing* (pp. 507–524). Marcel Dekker.

McNamara, T. (2013). Language testing: History, validity, policy. *APA handbook of testing and assessment in psychology, Vol. 1: Test theory and testing and assessment in industrial and organizational psychology.*, *1*, 341–352. Retrieved from http://content.apa.org/books/14047-021 doi: 10.1037/14047-021

Messick, S. (1960). Dimensions of social desirability. *Journal of Consulting Psychology*, *24*(4), 279–287. Retrieved from http://content.apa.org/journals/ccp/24/4/279 doi: 10.1037/h0044153

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11. Retrieved from http://edr.sagepub.com/content/18/2/5.abstract doi: 10.3102/0013189x018002005

Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, *50*(9), 741–749.

Messick, S., & Jackson, D. N. (1961). Desirability scale values and dispersons for MMPI Items. *Psychological Reports*, *8*, 409–414.

Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review*, *80*(4), 252–283. doi: 10.1037/h0035002

Mitra, S. K., & Fiske, D. W. (1956). Intra-individual variability as related to test score and item. *1Educational and Psychological Measurement*, *16*, 3–12.

Montague, R. (2008). The proper treatment of quantification in ordinary English. *Formal Semantics: The Essential Readings*, 17–34. doi: 10.1002/9780470758335.ch1

NSERC, C. S. (2014). *TCPS: Ethical Conduct for Research Involving Humans*. Retrieved from `http://www.pre.ethics.gc.ca/eng/policy-politique/initiatives/tcps2-eptc2/Default/` doi: 1

O'Grady, W., Archibald, J., Aronoff, M., & Rees-Miller, J. (2010). *Contemporary Linguistics: An Introduction* (6th ed.). Boston, MA: Bedford/St. Martin's.

Owens, W. A., Glennon, J. R., & Albright, L. E. (1962). Retest consistency and the writing of life history items: A first step. *Journal of Applied Psychology*, *46*(5), 329–331.

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

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press.

Payne, F. D. (1974). Relationships between response stability and item endorsement, social desirability, and ambiguity in the MMPI and the CPI. *Multi*, *3171*(March), 127–148. doi: 10.1207/s15327906mbr0902

Piccinelli, M., & Wilkinson, G. (1999). Gender differences in depression Critical review. *British Journal of Psychiatry*, 486–493.

Pinker, S. (1994). *The Language Instinct*. New York: HarperCollins.

Pinter, R., & Forlano, G. (1938). Four retests of a personality inventory. *Journal of Educational Psychology*, *29*, 93–100. doi: 10.1037/h0059933

Pratt, L. A., & Brody, D. J. (2014). Depression in the U.S. Household Population, 2009 âĂŞ 2012. *NCHS Data Brief*(172), 2009–2012.

Princeton University, N. (2010). *About WordNet.* Retrieved from `http://wordnet.princeton.edu`

Radloff, L. S. (1977). A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*(3), 385–401. doi: 10.1177/014662167700100306

Radloff, L. S. (1991). The use of the Center for Epidemiologic Studies Depression Scale in adolescents and young adults. *Journal of youth and adolescence*, *20*(2), 149–66. doi: 10.1007/BF01537606

Rodriguez, M. C., & Haladyna, T. M. (2013). Objective testing of educational achievement. *APA handbook of testing and assessment in psychology: Test theory and testing and assessment in industrial and organizational psychology*, *1*(Mc), 305–314. Retrieved from `http://content.apa.org/books/14047-018` doi: 10.1037/14047-018

Rorer, L. G. (1965). The great response-style myth. *Psychological Bulletin*, *63*(3), 129–156.

Roszkowski, M. J., & Soven, M. (2010). Shifting gears : consequences of including two negatively worded items in the middle of a positively worded questionnaire. *Assessment & Evaluation in Higher Education*, *35*(1), 117–134. doi: 10.1080/02602930802618344

Rotter, J. B. (1954). The clinical measurement of personality. In *Social learning and clinical psychology* (pp. 243–334). Englewood Cliffs, NJ: US: Prentice-Hall, Inc. doi: 10.1037/h0063470

Rumbaut, R. G. (1994). The crucible within: ethnic identity, self-esteem, and segmented assimilation among children of immigrants. *International Migration Review*, *28*(4), 748–794. doi: 10.1177/019027250707000107

Rush, T. R. (1984). Assessing readability: formulas and alternatives. In *Annual meeting of the wyoming state reading council of the international readaing association* (p. 19).

Santor, D. a., Gregus, M., & Welch, A. (2006). Eight decades of measurement in depression. *Measurement: Interdisciplinary Research & Perspective*, *4*(3), 135–155. doi: 10.1207/s15366359mea0403_1

Schalet, B. D., Cook, K. F., Choi, S. W., & Cella, D. (2014). Establishing a Common Metric for Depressive Symptoms: Linking the BDI-II, CES-D, and PHQ-9 to PROMIS Depression. *Psychological Assessment*, *26*(2), 88–513–527. Retrieved from `http://dx.doi.org/10.1016/j.janxdis.2013.11.006{%}5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/24548149{%}5Cnhttp://linkinghub.elsevier.com/retrieve/pii/S0887618513002156` doi: 10.1016/j.janxdis.2013.11.006

Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys*. Sage Publications.

Sebrechts, M. M., Enright, M., Bennet, R. E., & Martin, K. (1996). Using algebra word problems to assess quantitative ability: attributes, strategies, and errors. *Cognition and Instruction*, *14*(3), 285–343.

Shafer, A. B. (2006). Meta-analysis of the factor structures of four depressive symptomology questionnaires: Beck, CES-D, Hamilton, and Zung. *Journal of Clinical Psychology*, *62*(1), 123–146. doi: 10.1002/jclp

Sharp, K. L., Williams, A. J., Rhyner, K. T., & Ilardi, S. S. (2013). The clinical interview. *APA handbook of testing and assessment in psychology, Vol. 2: Testing and assessment in clinical and counseling psychology.*, *2*, 103–117. Retrieved from `http://content.apa.org/books/14048-007` doi: 10.1037/14048-007

Simpson, S. J. (1981). Meaning dominance and semantic context in the processing of lexical ambiguity. *Journal of Verbal Learning and Verbal Behavior*(20), 120–136.

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment*, *5*(4), 299–321.

Smith, G. T. (2005). On construct validity: Issues of method and measurement. *Psychological Assessment*, *17*(4), 396–408. Retrieved from `http://doi.apa.org/getdoi.cfm?doi=10.1037/1040-3590.17.4.396` doi: 10.1037/1040-3590.17.4.396

Smith, M. (1933). A note on stability in questionnaire response. *American Journal of Sociology*, *38*(5), 713–720.

Solís Salazar, M. (2015). The dilemma of combining positive and negative items in scales. *Psicothema*, *27*(2), 192–200. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/25927700 doi: 10.7334/psicothema2014.266

Sonnleitner, P. (2008). Using the LLTM to evaluate an item-generating system for reading comprehension. *Psychology Science Quarterly*, *50*(3), 345–362. Retrieved from http://p16277.typo3server.info/fileadmin/download/PschologyScience/3-2008/03{_}Sonnleitner.pdf

Stemler, S. E., & Sternberg, R. J. (2013). The assessment of aptitude. In *Apa handbook of testing and assessment in psychology: Vol. 3. testing and assessment in school psychology and education* (Vol. 3, pp. 281–296). doi: 10.1037/14049-013

Stricker, L. J. (1963). Acquiescence and social desirability response styles, item characteristics, and conformity. *Psychological Reports*(7438), 319–341.

Templeton, S., Cain, C. T., & Miller, J. O. (1981). Reconceptualizing readability: The relationship between surface and underlying structure analyses in predicting the difficulty of basal reader stories. *The Jounral of Educational Research*, *74*(6), 382–387.

Terborg, J. R., & Peters, L. H. (1974). Some observations on wording of item-stems for attitude questionnaires. *Psychological Reports*, *35*(7340), 463–466.

The Tsujii Laboratory, U. (2004). *Enju–A practical HPSG parser.* Retrieved from http://www.nactem.ac.uk/tsujii/software.html

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, *103*(3), 299–314.

Violato, C., & Harasym, P. H. (1987). Effects of structural characteristics of stem format of multiple-choice items on item difficulty and discrimination. *Psychological Reports*, *60*, 1259–1262.

Violato, C., & Marini, A. E. (1989). Effects of stem orientation and completeness of multiple-choice items on item difficulty and discrimination. *Educational and Psychological Measuremen*, *49*, 287–295. doi: 0803973233

Vonesh, E. F., & Chinchilli, V. M. (1997). *Linear and Nonlinear Models for the Analysis of Repated Measurements*. New York: Marcel Dekker.

Wason, P. C. (1961). Response to affirmative and negative binary statements. *British Journal of Psychology*, *52*(2), 133–142.

Weissmann, M. M., Sholomskas, D., Pottenger, M., Prusoff, B. A., & Locke, B. Z. (1977). Assessing depressive symptoms in five populations: A validation study. *Am. J. Epidemiol.*, *106*(3), 203–214. Retrieved from http://aje.oxfordjournals.org/content/106/3/203.short

Wendler, C., & Burrus, J. (2013). The importance of editorial reviews in ensuring item quality. *APA handbook of testing and assessment in psychology: Test theory and testing and assessment in industrial and organizational psychology*, *1*, 283–291. Retrieved from http://content.apa.org/books/14047-016 doi: 10.1037/14047-016

Werner, P. D., & Pervin, L. a. (1986). The content of personality inventory items. *Journal of personality and social psychology*, *51*(3), 622–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3761148 doi: 10.1037/0022-3514.51.3.622

Whitney, P. (1998). *The Psychology of Language*. Boston: Houghton Mifflin.

Wiggins, J. S. (1962). Strategic, method, and stylistic variance in the MMPI. *Psychological Bulletin*, *59*(3), 224–242. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-0011458719{&}partnerID=tZOtx3y1 doi: 10.1037/h0048092

Wiggins, J. S., & Goldberg, L. R. (1965). Interrelationships among MMPI item characteristics. *Educational and Psychological Measurement*, *XXV*(2), 381–397.

Wiggins, N. (1966). Individual viewpoints of social desirability. *Psychological Bulletin*, *66*(2), 68–77. Retrieved from http://psycnet.apa.org/journals/bul/66/2/68.pdf doi: 10.1037/h0023543

Winter, D., & Barenbaum, N. (1999). History of modern personality theory and research. In *Handbook of personality: Theory and research* (pp. 3–27). Retrieved from http://www.rc.usf.edu/{~}jdorio/Personality/HistoryofModernPersonalityTheoryandResearch.doc

Wood, A. M., Taylor, P. J., & Joseph, S. (2010). Does the CES-D measure a continuum from depression to happiness? Comparing substantive and artifactual models. *Psychiatry Research*, *177*(1-2), 120–123. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0165178110000491 doi: 10.1016/j.psychres.2010.02.003

Xerox. (2016). *Xerox Morphological Analysis.* Retrieved from https://open.xerox.com/Services/fst-nlp-tools/Consume/MorphologicalAnalysis-176

Zern, D. (1967). Effects of variations in question-phrasing on true-false answers by grade-school children. *Psychological Reports*, *20*, 527–533.

Zieky, M. J. (2013). Fairness review in assessment. *APA handbook of testing and assessment in psychology: Test theory and testing and assessment in industrial and organizational psychology*, *1*, 293–302. Retrieved from http://content.apa.org/books/14047-017 doi: 10.1037/14047-017

Appendix A

Focal Instrument

Here we list the 20 items for the Center for Epidemiological Studies Depression scale (CES-D). The response scale was: 0=Rarely (less than 1 day), 1=Some or a little of the time (1-2 days), 2=Occasionally (2-4 days), 3=Most of the time (5-7 days). Instructions said "As you read each statement, ask yourself how many times during the last week you felt that way."

Table A.1: CES-D Item Stems

1	I was bothered by things that usually don't bother me.
2	I did not feel like eating; my appetite was poor.
3	I felt I could not shake off the blues even with help from my friends.
4	I felt that I was just as good as other people.
5	I had trouble keeping my mind on what I was doing.
6	I felt depressed.
7	I felt everything I did was an effort.
8	I felt hopeful about the future.
9	I thought my life had been a failure.
10	I felt tearful.
11	My sleep was restless.
12	I was happy.
13	I talked less than usual.
14	I felt lonely.
15	People were unfriendly.
16	I enjoyed life.
17	I had crying spells.
18	I felt sad.
19	I felt that people dislike me.
20	I could not get going.

Appendix B

Linguistic Coding

B.1 Syntax Diagrams

B.1.1 Syntax Diagram Notation

To read the Syntax Diagrams:

CP: Complementizer Phrase

TP: Tense Phrase (i.e., where the tense for the verb is indicated)

VP: Verb Phrase

AP: Adjective Phrase

DP: Determiner Phrase

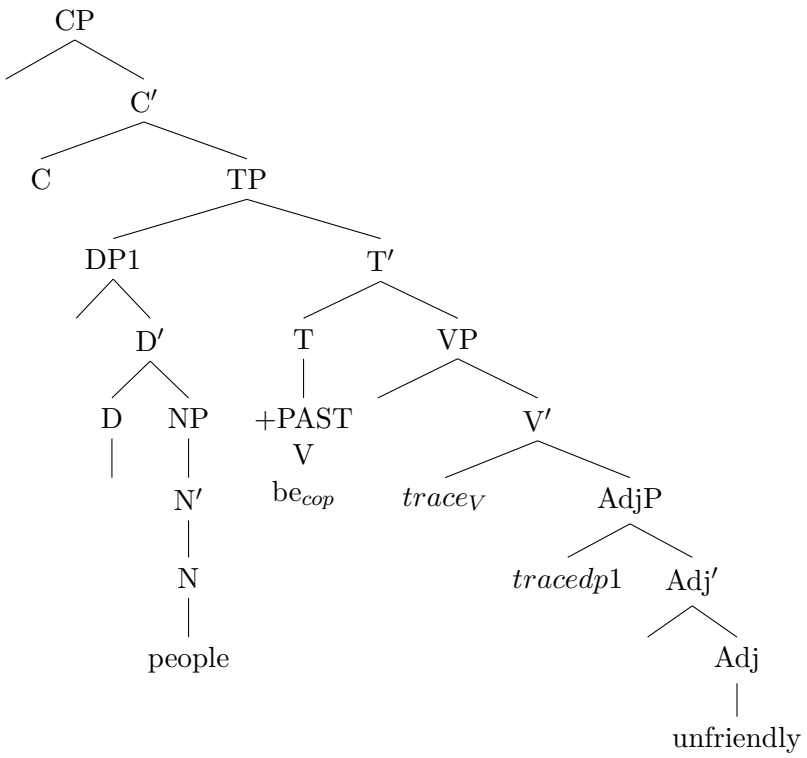
NP: Noun Phrase

Ø: Null determiner (i.e., for proper nouns or pronouns, you cannot have a determiner attach)

Trace: The trace is used to indicate where a particular phrase moved from to create the surface structure.

B.1.2 People were unfriendly.

Example Syntax Tree–Surface Structure


CP
C′
C
TP
DP1
T′
D′
T
VP
D
NP
+PAST
V
V′
N′
be$_{cop}$
$trace_V$
AdjP
N
$tracedp1$
Adj′
people
Adj
unfriendly

Appendix C

Correlational Analysis and Descriptive Statistics

Table C.1: Descriptive Statistics for Length Features

	Minimum	Maximum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
WC_nw	3.00	16.00	6.30	.84	3.77	14.22
WL_sym	1.00	2.00	1.30	.04	.20	.04
WL_syd	.00	1.00	.52	.04	.20	.04
WL_ltm	2.67	6.67	3.93	.18	.80	.64
WL_ltd	1.53	4.04	2.36	.14	.64	.40

WC=Word Count, WL=Word Length, nw=Number of Words, sy=Syllables, lt=Letters, m=mean, d=Standard Deviation

Table C.2: Correlations for Length Features

	WC_nw	WL_sym	WL_syd	WL_ltm	WL_ltd
WC_nw	1				
WL_sym	-.30	1			
WL_syd	-.18	.87**	1		
WL_ltm	-.23	.85**	.71**	1	
WL_ltd	-.45*	.50*	.59**	.59**	1

[1] *. Correlation is significant at the 0.05 level (2-tailed). **. Correlation is significant at the 0.01 level (2-tailed).
[2] WC=Word Count, WL=Word Length, nw=Number of Words, sy=Syllables, lt=Letters, m=mean, d=Standard Deviation

Table C.3: Descriptive Statistics for Morphological Features, POS Count Variables

	Minimum	Maximum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
N	.00	3.00	.90	.19	.85	.73
V_Main	.00	3.00	1.20	.17	.77	.59
V_Cop	.00	1.00	.35	.11	.49	.24
V_Aux	.00	2.00	.25	.12	.55	.30
V_Mod	.00	1.00	.10	.07	.31	.09
Verb	1.00	4.00	1.90	.25	1.12	1.25
Adv	.00	2.00	.45	.17	.76	.58
Adj	.00	2.00	.65	.13	.59	.34
Comp	.00	1.00	.15	.08	.37	.13
Pro	.00	4.00	1.60	.22	.99	.99
Prep	.00	3.00	.45	.17	.76	.58
Det	.00	1.00	.20	.09	.41	.17

These measures consist of the total count for each item.
N=Noun; V_Main=Main Verb; V_Cop=Copular Verb; V_Aux=Auxiliary Verb; V_Mod=Modular Verb; Verb=Any type of Verb; Adv=Adverb; Adj=Adjective; Comp=Complementizer; Pro=Pronoun; Prep=Preposition; Det=Determiner

Table C.4: Descriptive Statistics for Morphological Features, POS proportion Variables

	Minimum	Maximum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
N	.00	.33	.13	.03	.12	.01
V_Main	.00	.40	.21	.03	.13	.02
V_Cop	.00	.33	.07	.03	.11	.01
V_Aux	.00	.18	.02	.01	.05	.00
V_Mod	.00	.20	.01	.01	.05	.00
Verb	.17	.60	.31	.02	.10	.01
Adv	.00	.20	.05	.02	.08	.01
Adj	.00	.33	.16	.03	.15	.02
Comp	.00	.17	.02	.01	.04	.00
Pro	.00	.38	.26	.02	.09	.01
Prep	.00	.20	.05	.02	.07	.00
Det	.00	.17	.02	.01	.05	.00

These measures consist of the total count for each POS divided by the total number of words for the item.
N=Noun; V_Main=Main Verb; V_Cop=Copular Verb; V_Aux=Auxiliary Verb; V_Mod=Modular Verb; Verb=Any type of Verb; Adv=Adverb; Adj=Adjective; Comp=Complementizer; Pro=Pronoun; Prep=Preposition; Det=Determiner

Table C.5: Correlations for Morphological Features, POS, Count

	N	Main	Cop	Aux	Mod	Vrb	Adv	Adj	Cmp	Pro	Prp
N	1										
Main	.35	1									
Cop	.21	-.48*	1								
Aux	.39	.37	.05	1							
Mod	.24	.36	-.24	-.16	1						
Vrb	.60**	.76**	.07	.73**	.34	1					
Adv	.32	.29	-.02	.35	.47*	.49*	1				
Adj	-.39	-.65**	.27	-.37	-.38	-.62**	.02	1			
Cmp	.39	.26	-.01	-.20	.33	.17	.50*	.01	1		
Pro	.63**	.80**	-.02	.58**	.14	.86**	.39	-.52*	.32	1	
Prp	.64**	.38	-.16	.22	.47*	.43	.73**	-.10	.50*	.53*	1
Det	.51*	.20	.16	.00	.25	.28	.03	-.35	.14	.34	.37

These measures consist of the total count for each item.
N=Noun; Main=Main Verb; Cop=Copular Verb; Aux=Auxiliary Verb; Mod=Modular Verb; Vrb=Any type of Verb; Adv=Adverb; Adj=Adjective; Cmp=Complementizer; Pro=Pronoun; Prp=Preposition; Det=Determiner

Table C.6: Correlations for Morphological Features, POS Ratio

	N	Main	Cop	Aux	Mod	Vrb	Adv	Adj	Cmp	Pro	Prp
N	1										
Main	-.38	1									
Cop	.28	-.79**	1								
Aux	.16	-.15	-.07	1							
Mod	-.22	.30	-.18	-.14	1						
Vrb	-.20	.46*	-.01	.20	.59**	1					
Adv	-.30	-.05	-.24	.23	.49*	.01	1				
Adj	-.35	-.15	.31	-.46*	-.32	-.25	-.32	1			
Cmp	.06	.05	-.15	-.18	-.01	-.21	.13	-.28	1		
Pro	-.37	.55*	-.34	.04	-.21	.25	-.37	-.09	.00	1	
Prp	-.04	-.25	-.29	.16	-.01	-.59**	.56**	-.26	.09	-.37	1
Det	.20	-.14	-.01	.07	-.05	-.18	-.20	-.33	-.09	-.08	.25

[1] *. Correlation is significant at the .05 level (2-tailed). **. Correlation is significant at the .01 level (2-tailed).
N=Noun; Main=Main Verb; Cop=Copular Verb; Aux=Auxiliary Verb; Mod=Modular Verb; Vrb=Any type of Verb; Adv=Adverb; Adj=Adjective; Cmp=Complementizer; Pro=Pronoun; Prp=Preposition; Det=Determiner

Table C.7: Descriptive Statistics for Morphological Features, Morphemes, Count Variables

	Minimum	Maximum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
Content	2.00	7.00	3.50	.36	1.61	2.58
Function	.00	9.00	2.80	.53	2.38	5.64
Derv	.00	1	.65	.11	.49	.24
Bound	.00	3.00	1.10	.18	.79	.62
Infl	1	4.00	1.65	.22	.99	.98
Suff	.00	3.00	.80	.17	.77	.59
Pref	.00	1	.15	.08	.37	.13

These measures consist of the total count for each item.
Derv=Derivation; Bound=Bound Morpheme; Infl=Inflection; Suff=Suffix; Pref=Prefix

Table C.8: Descriptive Statistics for Morphological Features, Morphemes, Proportion Variables

	Minimum	Maximum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
Content	.36	1	.61	.03	.14	.02
Function	.00	.64	.39	.03	.14	.02
Derv	.00	.33	.14	.03	.13	.02
Bound	.00	.67	.21	.04	.17	.03
Infl	.17	.38	.28	.02	.08	.01
Suff	.00	.33	.15	.03	.13	.02
Pref	.00	.33	.03	.02	.08	.01

These measures consist of the total count for each item.
Derv=Derivation; Bound=Bound Morpheme; Infl=Inflection; Suff=Suffix; Pref=Prefix

Table C.9: Correlations for Morphological Features, Morphology, Count

	Cont	Func	Infl	Derv	Bound	Pref	Suff
Cont	1						
Func	.79**	1					
Infl	.68**	.84**	1				
Derv	-.23	-.20	-.16	1			
Bound	.17	.40	.32	.23	1		
Pref	-.04	-.08	.01	.31	-.05	1	
Suff	.21	.47*	.32	.50*	.64**	-.07	1

[1] *. Correlation is significant at the .05 level (2-tailed). **. Correlation is significant at the .01 level (2-tailed).
[2] Cont=Content; Func= function; Derv=Derivation; Bound=Bound Morpheme; Infl=Inflection; Suff=Suffix; Pref=Prefix

Table C.10: Correlations for Morphological Features, Morphology, Ratio

	Cont	Func	Infl	Derv	Bound	Pref	Suff
Cont	1						
Func	-1**	1					
Infl	.19	-.19	1				
Derv	.49*	-.49*	.15	1			
Bound	.59**	-.59**	.15	.67**	1		
Pref	.44	-.44	.11	.32	.42	1	
Suff	.34	-.34	.18	.88**	.68**	.16	1

[1] *. Correlation is significant at the .05 level (2-tailed). **. Correlation is significant at the .01 level (2-tailed).
[2] Cont=Content; Func= function; Derv=Derivation; Bound=Bound Morpheme; Infl=Inflection; Suff=Suffix; Pref=Prefix

Table C.11: Morphological Features (POS) Correlated with Word Count

	N	Main	Cop	Aux	Mod	Verb	Adv	Adj	Cmp	Pro	Prp	Det
WC_A	.80**	.60**	.08	.49*	.38	.79**	.72**	-.31	.54*	.83**	.83**	.44
WC_B	.17	-.24	-.19	.49*	.10	-.23	.48*	-.65**	.32	-.14	.62**	.26

[1] * Correlation is significant at the .05 level (2-tailed).
[2] ** Correlation is signifcant at the .01 level (2-tailed)
[A] POS counts
[B] POS proportions
N=Noun; Main=Main Verb; Cop=Copular Verb; Aux=Auxiliary Verb; Mod=Modular Verb; Verb=Any type of Verb; Adv=Adverb; Adj=Adjective; Cmp=Complementizer; Pro=Pronoun; Prp=Preposition; Det=Determiner

Table C.12: Morphological Features Correlated with Word Count

	Content	Function	Infl	Derv	Bound	Pref	Suff
Word Count A	.92**	.96**	.82**	-.23	.33	-.07	.39
Word Count B	-.68**	.68**	-.33	-.50*	-.41	-.16	-.27

[1] * Correlation is significant at the .05 level (2-tailed). [2] ** Correlation is significant at the .01 level (2-tailed)
Infl=Inflection; Derv=Derivation; Bound=Bound Morpheme; Pref=Prefix; Suff=Suffix

Table C.13: Descriptive Statistics for Syntactic Features

	Minmum	Maximum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
NP	.00	1	.23	.07	.30	.09
Tense_Pst	1	1	1	.00	.00	.00
Tense_Pres	.00	1	.10	.07	.31	.09
Voice_Act	1	1	1	.00	.00	.00
Voice_Pass	.00	1	.05	.05	.22	.05
Aspect_Prog	.00	1	.10	.07	.31	.09
Aspect_Perf	.00	1	.05	.05	.22	.05
Ambig	.00	1	.30	.11	.47	.22
Neg	.00	1	.30	.11	.47	.22
Idiom	.00	1	.10	.07	.31	.09
Coll	.00	1	.25	.10	.44	.20
Trace	.00	1	.11	.07	.32	.10
Rel	.00	1	.15	.08	.37	.13
Poss	.00	1	.25	.10	.44	.20

NP=Noun Phrase; Tense_pst=Past Tense; Tense_pres=Present tense; Voice_act=Active voice; Voice_Pass=Passive Voice; Aspect_Prog=Progressive; Aspect_Perf=Perfect; Ambig=Ambiguity; Neg=Negation; Idiom=Idiomatic; Coll=Collocation; Trace=Trace; Rel=Relative Clause; Poss=Possessive

Table C.14: Syntactic Correlations

	Np	Pres	Pass	Prog	Perf	Ambig	Neg	Idiom	Coll	Trace	Rel	Poss
Np	1											
Pres	-.27	1										
Pass	-.18	.69**	1									
Prog	-.12	-.11	-.08	1								
Perf	.34	-.08	-.05	-.08	1							
Ambig	.13	-.22	-.15	.15	-.15	1						
Neg	-.19	.51*	.35	.15	-.15	.29	1					
Idiom	.25	-.11	-.08	-.11	-.08	.51*	.51*	1				
Coll	.09	-.19	-.13	.19	-.13	.88**	.38	.58**	1			
Trace	.17	-.12	-.08	.44	-.08	-.20	-.23	-.12	-.20	1		
Rel	.03	.33	.55*	.33	-.10	-.28	.03	-.14	-.24	.79**	1	
Poss	.66**	-.19	-.13	.19	.40	.13	.13	.58**	.20	.18	.08	1

[1] * Correlation is significant at the .05 level (2-tailed). [2] ** Correlation is signifcant at the .01 level (2-tailed)

NP= Noun Phrase; Tense_pst=Past Tense; Tense_pres=Present tense; Voice_act=Active voice; Voice_Pass=Passive Voice; Aspect_Prog=Progressive; Aspect_Perf=Perfect; Ambig=Ambiguity; Neg=Negation; Idiom=Idiomatic; Coll=Collocation; Trace=Trace; Rel=Relative Clause; Poss=Possessive

Table C.15: Semantic Descriptives

	Minimum	Maximum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
Freq_con	1.62	3.51	2.64	.12	.52	.27
Freq_all	1.62	3.53	2.84	.12	.53	.28
Freq_min	.00	3.37	1.19	.22	.96	.93
Fam	561.00	620.50	591.30	3.32	14.85	220.40
Concrete	281.67	432.00	328.50	8.06	36.03	1298.00
Image	293.67	482.00	382.24	10.29	46.01	2117.06
Meaning	331.00	557.00	456.55	14.72	65.83	4333.17
Poly	3.14	11.83	5.24	.44	1.96	3.86
Hyper_n	.00	8.00	4.03	.70	3.15	9.92
Hyper_v	.73	4.17	1.45	.17	.74	.55
Hyper_nv	.21	1.85	1.06	.12	.54	.29
Poly_FN	1.25	7.00	3.15	.36	1.62	2.63
Poly_WN	3.00	22.33	8.54	.93	4.18	17.47
Spec	.00	1.00	.25	.10	.44	.20
Concrete	.00	1.00	.65	.11	.49	.24
Culture	.00	1.00	.45	.11	.51	.26
Gender	.00	1.00	.35	.11	.49	.24
Pos_af	.00	1.00	.20	.09	.41	.17
Somatic	.00	1.00	.35	.11	.49	.24
Dep_af	.00	1.00	.35	.11	.49	.24
Inter	.00	1.00	.10	.07	.31	.09
Ref	.00	1.00	.20	.09	.41	.17
Behav	.00	1.00	.30	.11	.47	.22
Feel	.00	1.00	.70	.11	.47	.22

[1] Freq_con: Logarithm of averaged CELEX frequencies for content words; Freq_all: Logarithm of averaged CELEX frequencies for all words; Freq_all: Minimum logarithm of averaged CELEX frequencies for content words.
[2] Fam: Average familiarity; Concrete: Average concreteness rating; Image: Average Imageability; Meaning: Average Meaningfulness.
[3] Poly: Average polysymy for content words; Hyper_n: Average hypernymy for nouns; Hyper_v: Average hypernymy for verbs; Hyper_nv: Average hypernymy for nouns and verbs.
[4] Poly_FN: Average polysymy based on FrameNet; Poly_WN: Average polysymy based on WordNet; Spec: Specificity, manual rating; Concrete: Concreteness, manual rating; Culture: Cultural loading, manual rating; Gender: Gender loading, manual rating.
[5] Pos_af: Positive affect subscale, reverse coded items; Somatic: Somatic subscale; Dep_af: Depressed affect subscale; Inter: Interpersonal problems subscale.
[6] Ref: Refers to others, manual; Behav: Describes a behavior, manual; Feel: Describes a feeling, manual.

Table C.16: Correlations for Semantic Features

	Fq_c	Fq_a	Fq_mc	Fam	Cnc	Img	Mea	Pol	Hy1	Hy2	Hy3	fn	wn	Spec	Conc	Cul	Gen	Pos	Som	Dep	Int	Ref	Beh
Freq_c	1																						
Freq_a	.86**	1																					
Freq_mc	.30	.16	1																				
Fam	.30	.25	.28	1																			
Cnc	-.37	-.36	.02	.30	1																		
Img	-.39	-.38	.02	.21	.73**	1																	
Mea	-.64**	-.56*	.04	.27	.54*	.70**	1																
Pol	.58**	.42	.48*	.01	-.43	-.57**	-.62**	1															
Hyp1	.20	.46*	-.34	-.18	.11	.07	-.31	-.10	1														
Hyp2	-.24	-.17	.09	-.07	-.10	-.09	.27	-.17	-.30	1													
Hyp3	.23	.40	-.14	-.05	.11	.10	-.32	.16	.84**	-.18	1												
FN	-.03	-.23	.02	-.49*	-.34	-.29	-.33	.44	-.05	.03	.20	1											
WN	.59**	.51*	.47*	.10	-.39	-.51*	-.55*	.92**	-.05	-.21	.19	.39	1										
Spec	.08	.03	.10	-.30	-.01	-.17	-.35	-.05	.33	.27	.28	.22	-.16	1									
Conc	.28	.18	.13	.09	-.34	-.42	-.40	.38	-.06	.12	.13	.17	.14	.42	1								
Cul	-.01	-.16	-.37	.05	.13	-.03	.12	-.06	-.08	.03	.00	.28	-.05	-.06	.03	1							
Gen	-.27	-.15	-.41	-.37	-.23	-.31	.00	-.24	.15	.40	-.10	.15	-.33	.30	.10	.18	1						
Pos	.18	.19	.07	.27	-.15	.09	.11	-.15	.07	-.23	-.12	-.23	-.03	-.29	-.42	-.20	.16	1					
Som	.27	.35	.20	-.04	-.29	-.44	-.34	.27	.11	.39	.18	-.04	.24	.54*	.54*	-.24	.12	-.37	1				
Dep	-.41	-.50*	-.20	-.52*	-.05	.13	.06	-.04	-.25	-.01	-.15	.47*	-.14	-.18	-.12	.18	-.10	-.37	-.54*	1			
Int	-.02	-.02	-.10	.53*	.75**	.38	.29	-.16	.13	-.31	.11	-.38	-.11	-.19	-.10	.37	-.24	-.17	-.24	-.24	1		
Ref	.13	.18	-.39	.38	.53*	.25	.09	-.24	.27	-.18	.11	-.40	-.15	-.29	-.42	.30	-.10	.06	-.37	-.10	.67**	1	
Beh	.26	.17	.34	-.25	-.10	-.38	-.53*	.33	.16	.29	.23	.28	.21	.88**	.48*	-.15	.21	-.33	.66**	-.25	-.22	-.33	1

[1] Freq_c: Logarithm of averaged CELEX frequencies for content words; Freq_a: Logarithm of averaged CELEX frequencies for all words; Freq_mc: Minimum logarithm of averaged CELEX frequencies for content words.
[2] Fam: Average familiarity; Cnc: Average concreteness rating; Img: Average Imageability; Mea: Average Meaningfulness.
[3] Pol: Average polysymy for content words; Hyp1: Average hypernymy for nouns; Hyp2: Average hypernymy for verbs; Hyp3: Average hypernymy for nouns and verbs.
[4] FN: Average polysymy based on FrameNet; WN: Average polysymy based on WordNet; Spec: Specificity, manual rating; Conc: Concreteness, manual rating; Cul: Cultural loading, manual rating; Gen: Gender loading, manual rating.
[5] Pos: Positive affect subscale, reverse coded items; Som: Somatic subscale; Dep: Depressed affect subscale; Int: Interpersonal problems subscale.
[6] Ref: Refers to others, manual; Beh: Describes a behavior, manual;
[7] *. Correlation is significant at the 0.05 level (2-tailed). **. Correlation is significant at the 0.01 level (2-tailed).

Table C.17: Missing Data Summary, by Age, Sex, Ethnicity, and English Fluency

	Age	Sex	Ethnicity	English Fluency
Valid	2179	2180	2190	2184
Missing	22	21	11	17

Table C.18: Descriptive Statistics for Age, CES-D Composites, and English Fluency

	n	Min	Max	Mean	SE	Median	SD	σ^2	Skew	Kurt.
Age	2179	17	53	19.8	0.06	19	2.88	8.32	4.5	34.15
CES-D	2201	0	51	15.17	0.19	14	9.12	83.08	0.9	0.60
English Fl.	2184	1	4	1.8	0.02	1	1.02	1.05	1.0	-0.35
Valid N	2165									

Table C.19: Frequencies

	Frequency	Percent	Valid Percent	Cum. Percent
Sex				
Male	749	34.0	34.4	34.4
Female	1430	65.0	65.6	100.0
No Answer	1	.0	.0	100.0
Ethnicity				
African/African American/Carribean	31	1.4	1.4	1.4
Asian/Asian American/Pacific Islander	1129	51.3	51.6	53.0
European/American/Caucasian	735	33.4	33.6	86.5
Hispanic/Latinoa	19	.9	.9	87.4
First Nations/Native American	6	.3	.3	87.7
Metis	4	.2	.2	87.9
Biracial Non-Metis	30	1.4	1.4	89.2
Multi-racial Non-Metis	37	1.7	1.7	90.9
Other	189	8.6	8.6	99.5
No Answer	10	.5	.5	100.0
English Fluency				
VF	1174	53.3	53.8	53.8
MF	491	22.3	22.5	76.2
SF	294	13.4	13.5	89.7
LF	225	10.2	10.3	100.0

[1] VF: Very Fluent English is my first language;
MF: More fluent in English than my first language;
SF: Same fluency in English and my first language;
LF: Less fluent in English than in my first language.

Table C.20: Descriptive Statistics for Age, CES-D, by Sex

		n	Min	Max	Mean	SE	Median	SD	σ^2	Skew	Kurt.
Male											
	Age	743	17	53	19.95	.11	19	2.96	8.76	.09	.18
	CES-D	749	0	46	14.21	.31	13	8.41	70.78	.09	.18
Female											
	Age	1418	17	53	19.70	.07	19	2.81	7.90	.06	.13
	CES-D	1430	0	51	15.66	.25	15	9.44	89.09	.06	.13

Table C.21: Descriptive Statistics for Age, CES-D, by English Fluency

		n	Min	Max	Mean	SE	Med	SD	σ^2	Skew	Kurt.
VF											
	Age	1162	17	53	19.66	.08	19	2.85	8.11	.07	.14
	CES-D	1174	0	48	14.32	.27	13	9.14	83.49	.07	.14
MF											
	Age	486	17	53	19.36	.12	19	2.74	7.49	.11	.22
	CES-D	491	0	43	15.33	.39	14	8.53	72.79	.11	.22
SF											
	Age	293	17	33	19.74	.13	19	2.21	4.88	.14	.28
	CES-D	294	1	50	16.11	.54	15	9.23	85.17	.14	.28
LF											
	Age	224	17	42	21.44	.23	19	3.49	12.18	.16	.32
	CES-D	225	1	48	18.04	.63	15	9.49	89.99	.16	.32

[1] VF: Very Fluent English is my first language;
MF: More fluent in English than my first language;
SF: Same fluency in English and my first language;
LF: Less fluent in English than in my first language.

Appendix D

Linear Mixed Effects Models

D.1 Length

Table D.1: Length Features – Single Predictor Model Estimates

Feature	Parameter	Estimate	SE	df	t	p
Word Count	Intercept	-.097	.013	3191.5	-7.50	p<.006
	WC_nw	.005	.001	11169.8	5.31	p<.006
Avg. word length						
	Intercept	.272	.020	6987.6	13.32	p<.006
	WL_ltm	-.063	.004	700.9	-16.71	p<.006

WC=Word Count, WL=Word Length, nw=Number of Words, lt=Letters, m=mean
Avg word length: average number of letters per word.
Error rate is set at p=.00625

Table D.2: Length Features – Three Predictor Model Estimates

Feature	Parameter	Estimate	SE	df	t	p
Number of words	Intercept	.539	.013	14039.2	41.08	p<.006
	WC_nw	.011	.001	1010.2	8.76	p<.006
	Sex=Male	-.081	.011	11229.2	-7.67	p<.006
	Engfl	.051	.005	11227.6	1.45	p<.006
Avg. word length	Intercept	1.144	.028	6479.1	41.08	p<.006
	WL_ltm	-.118	.005	21095.5	-25.34	p<.006
	Sex=Male	-.069	.020	2166.3	-3.36	p<.006
	Engfl	.057	.009	2166.2	6.06	p<.006

WC=Word Count, WL=Word Length, nw=Number of Words, lt=Letters, m=mean
Avg word length: average number of letters per word.
Error rate is set at p=.00625

Table D.3: Length Features – English Fluency by Feature Interaction Models

Feature	Parameter	Estimate	SE	df	t	p
Number of Words	Intercept	.571	.024	364.0	23.62	p<.006
	WC_nw	.017	.002	546.0	8.67	p<.006
	Sex=Male	-.068	.020	2166.7	-3.33	p<.006
	Engfl	.062	.011	3813.5	5.57	p<.006
	Engfl * WC_nw	-.001	.001	5453.6	-.61	.543
Avg. word length	Intercept	.175	.042	6879.6	4.16	p<.006
	WL_ltm	-.038	.008	6923.6	-4.92	p<.006
	Sex=Male	-.183	.025	2277.4	-7.29	p<.006
	Engfl	.088	.020	7169.3	4.44	p<.006
	Engfl * WL_ltm	-.015	.004	6923.5	-3.97	p<.006

WC=Word Count, WL=Word Length, nw=Number of Words, lt=Letters, m=mean
Avg word length: average number of letters per word.
Error rate is set at p=.00625

Table D.4: Length Features – Sex by Feature Interaction Models

Feature	Parameter	Estimate	SE	df	t	p
Number of Words	Intercept	.576	.022	2787.4	25.90	p<.006
	WC_nw	.017	.001	41071.8	13.65	p<.006
	Sex=Male	-.081	.024	4272.2	-3.32	p<.006
	Engfl	.057	.009	2168.8	5.98	p<.006
	Sex=Male * WC_nw	.001	.002	41072.3	.71	.48
Avg. word length	Intercept	.323	.033	4485.0	9.84	p<.006
	WL_ltm	-.071	.005	6951.1	-15.24	p<.006
	sex=Male	-.280	.043	7057.1	-6.54	p<.006
	Engfl	.023	.012	2276.8	2.01	.045
	Sex=Male * WL_ltm	.022	.008	6949.1	2.74	.006

WC=Word Count, WL=Word Length, nw=Number of Words, lt=Letters, m=mean
Avg word length: average number of letters per word.
Error rate is set at p=.00625

Table D.5: Average Word Length, Letters – 5 predictor model

Feature	Parameter	Estimate	SE	df	t	p
Avg Word Length	Intercept	.207	.044	7076.7	4.72	p<.006
	WL_ltm	-.045	.008	6951.0	-5.49	p<.006
	Sex=Male	-.278	.043	7077.6	-6.49	p<.006
	Engfl	.087	.020	7083.7	4.40	p<.006
	Engfl * WL_ltm	-.015	.004	6950.1	-3.95	p<.006
	Sex=Male * WL_ltm	.022	.008	6947.8	2.71	.007

WC=Word Count, WL=Word Length, nw=Number of Words, lt=Letters, m=mean
Avg word length: average number of letters per word.
Error rate is set at p=.00625

D.2 Morphology

Table D.6: Morphological Features, Parts of Speech – One Predictor Model Estimates

Feature (proportions)	Parameter	Estimate	SE	df	t	p
Nouns	Intercept	.333	.012	1952.8	27.61	p<.001
	N	-.768	.027	8357.3	-28.47	p<.001
Main Verbs	Intercept	-.112	.013	2848.5	-8.45	p<.001
	V_Main	.191	.026	25661.9	7.39	p<.001
Adverbs	Intercept	-.013	.012	2293.6	-1.09	.277
	Adv	.590	.043	13302.9	13.62	p<.001
Pronouns	Intercept	-.159	.014	2823.1	-11.08	p<.001
	Pro	.616	.035	10901.4	17.59	p<.001
Adjectives	Intercept	-.016	.013	2499.4	-1.27	.205
	Adj	-.110	.021	9488.6	-5.12	p<.001
Aux Verbs	Intercept	-.035	.012	2483.4	-2.85	.004
	Aux	-.155	.060	7856.4	-2.56	.01
Copular Verbs	Intercept	-.003	.012	2435.0	-.25	.806
	Cop	-.213	.028	7932.1	-7.74	p<.001
Prepositions	Intercept	.042	.012	2146.2	3.54	p<.001
	Prep	.754	.051	7364.1	14.85	p<.001

Proportion=count of POS/Total words

Note that because the features were proportions, the coefficients are interpreted in terms of a 1 unit change, which is theoretically impossible. Thus, to be in line with the results reported in the text, the coefficient estimates would be divided by 10 to represent a .1 increase in the proportions.

Table D.7: Morphological Features, Parts of Speech – Three Predictor Model Estimates

Feature (proportions)	Parameter	Estimate	SE	df	t	p
Nouns	Intercept	.754	.021	2331.1	35.21	p<.001
	N	-.556	.031	8339.7	-17.75	p<.001
	Engfl	.057	.010	2164.3	5.98	p<.001
	Sex=Male	-.073	.021	2164.4	-3.55	p<.001
Main Verbs	Intercept	.643	.022	2549.7	29.56	p<.001
	V_Main	.181	.029	41071.9	6.22	p<.001
	Engfl	.057	.009	2168.8	5.99	p<.001
	Sex=Male	-.071	.020	2168.9	-3.50	p<.001
Adverbs	Intercept	.682	.021	2209.4	32.26	p<.001
	Adv	-.026	.044	14109.4	-.60	.55
	Engfl	.057	.010	2164.7	6.00	p<.001
	Sex=Male	-.072	.021	2164.8	-3.49	p<.001
Prounouns	Intercept	-.146	.027	2735.6	-5.38	p<.001
	Pro	.628	.035	10734.1	17.91	p<.001
	Engfl	.024	.012	2373.5	2.05	.04
	Sex=Male	-.189	.025	2373.6	-7.48	p<.001
Adjectives	Intercept	.004	.027	2454.2	.14	.89
	Adj	-.101	.021	9470.6	-4.69	p<.001
	Engfl	.020	.012	2355.0	1.68	.09
	Sex=Male	-.179	.026	2355.5	-6.97	p<.001
Copular	Intercept	.698	.021	2208.1	32.98	p<.001
	Cop	-.262	.031	13160.3	-8.33	p<.001
	Engfl	.057	.010	2164.7	5.98	p<.001
	Sex=Male	-.074	.021	2164.7	-3.59	p<.001
Prepositions	Intercept	.049	.025	2117.2	1.96	.05
	Prep	.750	.051	7325.9	14.74	p<.001
	Engfl	.026	.011	2111.8	2.23	.03
	Sex=Male	-.167	.025	2112.0	-6.78	p<.001

Proportion=count of POS/Total words

Note that because the features were proportions, the coefficients are interpreted in terms of a 1 unit change, which is theoretically impossible. Thus, to be in line with the results reported in the text, the coefficient estimates would be divided by 10 to represent a .1 increase in the proportions.

Error rate is set at .00156

Table D.8: Morphological Features, Parts of Speech – English Fluency by Feature Interaction Models

Feature (proportions)	Parameter	Estimate	SE	df	t	p
Noun	Intercept	.231	.026	2181.3	8.76	p<.001
	N	-.602	.056	25225.3	-10.79	p<.001
	Sex=Male	-.164	.024	1964.8	-6.97	p<.001
	Engfl	.041	.012	2177.0	3.39	p<.001
	Engfl * N	-.067	.027	25223.3	-2.47	.01
Verbs	Intercept	-.100	.028	2868.7	-3.51	p<.001
	V_Main	.285	.052	25245.4	5.44	p<.001
	Sex=Male	-.189	.026	2573.0	-7.38	p<.001
	Engfl	.026	.013	2852.9	1.99	.05
	Engfl * V_Main	-.047	.025	25239.9	-1.86	.06
Adjectives	Intercept	.680	.014	15130.2	48.89	p<.001
	Adj	-.425	.053	26612.5	-8.02	p<.001
	Sex=Male	-.082	.011	11113.0	-7.70	p<.001
	Engfl	.045	.006	14816.3	6.92	p<.001
	Engfl * Adj	.035	.026	26613.9	1.36	.17
Copular	Intercept	.694	.021	2359.9	32.56	p<.001
	Cop	-.264	.066	24373.4	-3.99	p<.001
	Sex=Male	-.069	.020	2166.2	-3.39	p<.001
	Engfl	.054	.010	2385.4	5.59	p<.001
	Engfl * Cop	.044	.032	24345.5	1.39	.17
Prepositions	Intercept	.632	.021	2375.5	29.62	p<.001
	Prep	.953	.099	13636.8	9.63	p<.001
	Sex=Male	-.068	.020	2166.6	-3.35	p<.001
	Engfl	.059	.010	2403.5	6.04	p<.001
	Engfl * Prep	-.007	.048	13616.1	-.16	.88

Proportion=count of POS/Total words

Note that because the features were proportions, the coefficients are interpreted in terms of a 1 unit change, which is theoretically impossible. Thus, to be in line with the results reported in the text, the coefficient estimates would be divided by 10 to represent a .1 increase in the proportions.

error rate is set at .00156

Table D.9: Morphological Features, Parts of Speech – Sex by Feature Interaction Models

Feature (proportions)	Parameter	Estimate	SE	df	t	p
Nouns	Intercept	.733	.012	16666.2	59.54	p<.001
	N	-.408	.043	40368.6	-9.54	p<.001
	Sex=Male	-.071	.014	25892.0	-4.93	p<.001
	Engfl	.055	.005	11547.1	11.08	p<.001
	Sex=Male * N	.016	.073	40363.6	.22	.825
Main verbs	Intercept	.621	.022	2761.9	28.01	p<.001
	V_Main	.261	.036	14949.2	7.32	p<.001
	Sex=Male	-.033	.024	4109.5	-1.38	.168
	Engfl	.057	.009	2165.9	5.98	p<.001
	Sex=Male * V_Main	-.173	.061	14951.7	-2.83	.005
Adjectives	Intercept	.727	.021	2431.5	33.81	p<.001
	Adj	-.312	.032	19338.4	-9.67	p<.001
	Sex=Male	-.020	.022	2999.1	-.89	.375
	Engfl	.058	.009	2166.4	6.12	p<.001
	Sex=Male * Adj	-.309	.055	19358.1	-5.61	p<.001
Copular verbs	Intercept	.641	.011	12367.7	57.19	p<.001
	Cop	-.583	.039	9140.3	-14.77	p<.001
	Sex=Male	-.105	.012	13680.5	-8.95	p<.001
	Engfl	.051	.005	11101.8	10.36	p<.001
	Sex=Male * Cop	.362	.067	9148.3	5.37	p<.001
Prepositions	Intercept	.634	.021	2243.2	30.18	p<.001
	Prep	.893	.060	13632.7	14.83	p<.001
	Sex=Male	-.074	.021	2403.4	-3.56	p<.001
	Engfl	.058	.009	2166.5	6.16	0p<.001
	Sex=Male * Prep	.133	.103	13635.6	1.30	.195

Proportion=count of POS/Total words

Note that because the features were proportions, the coefficients are interpreted in terms of a 1 unit change, which is theoretically impossible. Thus, to be in line with the results reported in the text, the coefficient estimates would be divided by 10 to represent a .1 increase in the proportions.

Error rate set at p=.00156

Table D.10: Morphological Features II-1-Predictor models

Feature	Parameter	Estimate	SE	df	t	p
Bound Morphemes	Intercept	.149	.013	2636.7	11.09	p<.004
	WC_ Bound	-.087	.005	21150.4	-19.37	p<.004
Suffixes	Intercept	-.039	.013	2870.4	-3.02	p<.004
	WC_Suff	-.043	.005	13783.7	-9.30	p<.004
Content	Intercept	.196	.019	4780.2	10.17	p<.004
	WR_Content	-.327	.022	13340.7	-14.98	p<.004

WC=Word count; WR=Proportion
Error rate set to .004

Table D.11: Morphological Features II – 3 Predictor Models

Feature	Parameter	Estimate	SE	df	t	p
Bound Morphemes	Intercept	.155	.026	2370.6	5.95	p<.004
	WC_ Bound	-.088	.005	21070.4	-19.45	p<.004
	Engfl	.030	.011	2199.0	2.62	.009
	Sex=Male	-.162	.025	2199.8	-6.58	p<.004
Suffixes	Intercept	-.016	.027	2679.8	-.59	.554
	WC_Suff	-.042	.005	13792.4	-9.07	p<.004
	Engfl	.020	.012	2582.7	1.67	.096
	Sex=Male	-.174	.026	2582.6	-6.80	p<.004
Content	Intercept	.221	.030	3504.6	7.37	p<.004
	WR_content	-.327	.022	13406.6	-14.94	p<.004
	Engfl	.021	.012	2325.7	1.75	.081
	Sex=Male	-.174	.025	2326.3	-6.88	p<.004

WC=Word count; WR=Proportion
Error rate set to .004

Table D.12: Morphological Features II – English Fluency by Feature Interactions

Feature	Parameter	Estimate	SE	df	t	p
Bound Morphemes	Intercept	.630	.016	17877.0	40.29	p<.004
	WC_ Bound	-.032	.010	14814.6	-3.35	p<.004
	Engfl	.102	.007	17499.3	13.87	p<.004
	Sex=Male	-.081	.010	11153.2	-7.73	p<.004
	Engfl * WC_ Bound	-.042	.005	14816.0	-9.03	p<.004

WC=Word count; WR=Proportion
Error rate set to .004

Table D.13: Morphological Features II – Sex by Feature Interactions

Feature	Parameter	Estimate	SE	df	t	p
Bound Morphemes	Intercept	.134	.026	2451.6	5.07	p<.004
	WC_ Bound	-.075	.006	21103.4	-13.42	p<.004
	eng_fl	.030	.011	2206.5	2.65	.008
	sex=Male	-.105	.028	2601.5	-3.72	p<.004
	sex=Male * WC_ Bound	-.038	.009	21081.0	-4.02	p<.004

WC=Word count; WR=Proportion
Error rate set to .004

Table D.14: Morphological Features II – 5 Predictor Model

Feature	Parameter	Estimate	SE	df	t	p
Bound Morphemes	Intercept	.027	.029	2614.9	.94	.348
	WC_ Bound	-.005	.010	21106.2	-.51	.612
	Engfl	.089	.013	2613.5	6.80	.000
	Sex=Male	-.102	.028	2614.7	-3.62	.000
	Sex=Male*WC_ Bound	-.039	.009	21091.4	-4.12	.000
	Engfl * WC_ Bound	-.039	.004	21111.0	-8.90	.000

WC=Word count; WR=Proportion
Error rate set to .004

D.3 Syntax

Table D.15: Syntax – Single Predictor Models

Feature	Parameter	Estimate	SE	df	t	p
Ambiguity	Intercept	-.058	.012	2680.8	-4.67	p<.003
	Ambig	-.073	.007	21914.6	-10.42	p<.003
Negations	Intercept	-.014	.013	2509.3	-1.07	.285
	Neg	-.077	.007	32260.1	-11.15	p<.003
Relative Clause	Intercept	.258	.011	1337.6	24.06	p<.003
	Rel	.419	.010	6005.0	40.20	p<.003
Possessive	Intercept	-.062	.012	2665.5	-5.00	p<.003
	Poss	-.047	.008	15958.8	-6.22	p<.003

Error rate set to .003

Table D.16: Syntax – 3 Predictor Model Estimates

Feature	Parameter	Estimate	SE	df	t	p
Ambiguity	Intercept	.718	.021	2225.1	34.19	p<.003
	Ambig	-.138	.008	15112.2	-17.79	p<.003
	Sex=Male	-.069	.020	2166.1	-3.37	p<.003
	Engfl	.057	.009	2166.0	6.00	p<.003
Negations	Intercept	-.001	.026	2525.5	-.02	.981
	Neg	-.079	.007	31649.7	-11.52	p<.003
	Sex=Male	-.187	.026	2480.4	-7.29	p<.003
	Engfl	.026	.012	2480.1	2.17	.030
Possessives	Intercept	-.033	.026	2614.6	-1.25	.210
	Poss	-.049	.008	15452.7	-6.45	p<.003
	Sex=Male	-.185	.026	2587.6	-7.21	p<.003
	Engfl	.018	.012	2587.9	1.49	.138
Relative clauses	Intercept	.597	.021	2189.2	28.50	p<.003
	Rel	.538	.010	20223.3	54.26	p<.003
	Sex=Male	-.068	.020	2167.0	-3.31	p<.003
	Engfl	.059	.009	2167.0	6.20	p<.003

Error rate set to .003

Table D.17: Syntax – Feature by English Fluency Interaction Models

Feature	Parameter	Estimate	SE	df	t	p
Ambiguity	Intercept	.709	.021	2403.5	33.09	p<.003
	Ambig	-.095	.016	41071.4	-5.93	p<.003
	Sex=Male	-.071	.020	2168.9	-3.50	p<.003
	Engfl	.061	.010	2435.6	6.25	p<.003
	Engfl * Ambig	-.014	.008	41070.9	-1.81	.071
Relative clauses	Intercept	.257	.023	1274.8	11.18	p<.003
	Rel	.463	.021	5937.7	21.93	p<.003
	Sex=Male	-.141	.022	1298.3	-6.28	p<.003
	Engfl	.025	.010	1271.5	2.42	.016
	Engfl * Rel	-.025	.010	5935.7	-2.47	.014
Possessives	Intercept	-.041	.026	2650.5	-1.56	.119
	Poss	-.008	.015	15243.2	-.55	.580
	Sex=Male	-.186	.026	2567.6	-7.24	p<.003
	Engfl	.022	.012	2655.6	1.82	.069
	Engfl * Poss	-.022	.007	15244.1	-3.04	p<.003

Error rate set to .003

Table D.18: Syntax – Sex by Feature Interaction Models

Feature	Parameter	Estimate	SE	df	t	p
Ambiguity	Intercept	-.038	.026	2648.9	-1.43	.154
	Ambig	-.052	.009	22496.4	-6.00	p<.003
	Sex=Male	-.157	.026	2690.1	-6.05	p<.003
	Engfl	.020	.012	2603.1	1.68	.092
	Sex=Male * Ambig	-.061	.015	22476.1	-4.10	p<.003
Relative clause	Intercept	.597	.021	2200.8	28.46	p<.003
	Rel	.538	.012	20214.8	43.97	p<.003
	Sex=Male	-.068	.021	2270.0	-3.28	p<.003
	Engfl	.059	.009	2167.0	6.20	p<.003
	Sex=Male * Rel	.001	.021	20230.4	.04	.965
Possessives	Intercept	-.029	.026	2636.3	-1.12	.263
	Poss	-.068	.009	15648.0	-7.36	p<.003
	Sex=Male	-.198	.026	2686.9	-7.67	p<.003
	Engfl	.018	.012	2595.9	1.52	.128
	Sex=Male * Poss	.061	.016	15666.8	3.83	p<.003

Error rate set to .003

Table D.19: Syntax – 5 Predictor Model

Feature	Parameter	Estimate	SE	df	t	p
Possessives	Intercept	-.038	.026	2666.9	-1.44	.151
	Poss	-.029	.016	15440.4	-1.78	.075
	Sex=Male	-.199	.026	2666.6	-7.69	p<.003
	Engfl	.022	.012	2665.9	1.86	.063
	Sex=Male * Poss	.060	.016	15461.3	3.78	p<.003
	Engfl * Poss	-.022	.007	15447.5	-2.97	.003

Error rate set to .003

D.4 Semantics

Table D.20: Semantics I – Single Predictor Models

Feature	Parameter	Estimate	SE	df	t	p
Specificity	Intercept	-.068	.012	2661.2	-5.58	p<.001
	Spec	-.042	.008	14128.2	-5.56	p<.001
Concreteness	Intercept	-.020	.013	2760.4	-1.51	.132
	Conc	-.049	.007	20502.2	-7.10	p<.001
Culture	Intercept	.498	.011	2114.5	43.73	p<.001
	Cult	-.267	.007	33085.7	-39.87	p<.001
Gender	Intercept	-.066	.012	2687.0	-5.29	p<.001
	Gender	-.052	.007	25850.7	-7.73	p<.001
Positive Affect	Intercept	.110	.011	1927.3	9.71	p<.001
	Pos	.232	.009	9223.8	26.36	p<.001
Somatic Subscale	Intercept	.203	.011	1567.5	18.96	p<.001
	Som	.232	.007	14912.9	31.25	p<.001
Depressed Affect	Intercept	.198	.012	1834.1	15.88	p<.001
	Dep	-.160	.007	29502.0	-23.65	p<.001
Interpersonal Probls	Intercept	.211	.012	912.5	17.58	p<.001
	Int	-.196	.010	11903.8	-20.55	p<.001
Refers to others	Intercept	.014	.013	2422.2	1.09	.277
	Ref	-.105	.008	21160.9	-13.53	p<.001
Behavior/feeling	Intercept	-.080	.012	2660.6	-6.52	p<.001
	Beh	-.014	.007	16420.4	-2.01	.045

Error rate set to .001

Table D.21: Semantics I – Three Predictor Models

Feature	Parameter	Estimate	SE	df	t	p
Specificity	Intercept	-.042	.026	2606.0	-1.61	.109
	Spec	-.041	.008	14381.3	-5.38	p<.001
	Sex=Male	-.176	.026	2581.9	-6.88	p<.001
	Engfl	.019	.012	2581.8	1.64	.102
Concreteness	Intercept	.000	.027	2653.2	.02	.988
	Conc	-.048	.007	20091.1	-6.92	p<.001
	Sex=Male	-.174	.025	2494.7	-6.81	p<.001
	Engfl	.021	.012	2494.7	1.76	.078
Culture	Intercept	.488	.022	2051.0	21.97	p<.001
	Cult	-.268	.007	33080.3	-39.95	p<.001
	Sex=Male	-.132	.021	1978.1	-6.28	p<.001
	Engfl	.033	.010	1978.4	3.35	.001
Gender	Intercept	-.040	.026	2623.1	-1.52	.128
	Gender	-.048	.007	26774.1	-7.13	p<.001
	Sex=Male	-.168	.026	2587.6	-6.53	p<.001
	Engfl	.019	.012	2587.4	1.57	.116
Positive affect	Intercept	.087	.024	1906.8	3.58	p<.001
	Pos	.233	.009	9101.8	26.56	p<.001
	Sex=Male	-.143	.024	1909.5	-6.05	p<.001
	Engfl	.042	.011	1909.1	3.80	p<.001
Somatic subscale	Intercept	.215	.023	1506.4	9.35	.000
	Som	.231	.007	14697.0	31.13	p<.001
	Sex=Male	-.135	.022	1506.9	-6.03	p<.001
	Engfl	.021	.010	1506.5	1.98	.048
Depressed affect	Intercept	.213	.025	1714.0	8.51	p<.001
	Dep	-.160	.007	28673.5	-23.54	p<.001
	Sex=Male	-.137	.024	1674.5	-5.71	p<.001
	Engfl	.024	.011	1674.5	2.18	.030
Interpersonal problems	Intercept	.179	.025	1191.9	7.16	p<.001
	Int	-.191	.010	11593.6	-20.09	p<.001
	Sex=Male	-.170	.024	1177.4	-7.02	p<.001
	Engfl	.030	.011	1176.4	2.66	.008
Refers to others	Intercept	.032	.026	2417.3	1.23	.220
	Ref	-.107	.008	20305.8	-13.76	p<.001
	Sex=Male	-.187	.025	2374.9	-7.36	p<.001
	Engfl	.021	.012	2374.5	1.81	.070

Error rate set to .001

Table D.22: Semantics I – Feature by English Fluency Interaction Models

Feature	Parameter	Estimate	SE	df	t	p
Culture	Intercept	.815	.022	2617.8	37.24	p<.001
	Cult	-.299	.015	41070.9	-20.54	p<.001
	Sex=Male	-.071	.020	2168.8	-3.50	p<.001
	Engfl	.042	.010	2679.9	4.24	p<.001
	Cult * Engfl	.032	.007	41070.6	4.52	p<.001
Positive Affect	Intercept	.079	.024	1913.0	3.27	.001
	Pos	.024	.018	9476.2	1.39	.165
	Sex=Male	-.143	.024	1915.2	-6.02	p<.001
	Engfl	.043	.011	1912.2	3.94	p<.001
	Pos * Engfl	.118	.008	9474.0	13.87	p<.001
Interpersonal problems	Intercept	.142	.026	1181.3	5.49	p<.001
	Int	-.096	.019	11709.2	-4.98	p<.001
	Sex=Male	-.167	.024	1149.9	-6.91	p<.001
	Engfl	.052	.012	1179.8	4.38	p<.001
	Int * Engfl	-.053	.009	11709.6	-5.69	p<.001

Error rate set to .001

Table D.23: Semantics I – Sex by Feature Interaction Models

Feature	Parameter	Estimate	SE	df	t	p
Culture	Intercept	.484	.023	2076.7	21.50	p<.001
	Cult	-.263	.008	33104.9	-31.80	p<.001
	Sex=Male	-.122	.024	2083.6	-5.10	p<.001
	Engfl	.033	.010	1979.3	3.35	.001
	Sex=Male * Cult	-.014	.014	33080.5	-1.01	.315
Positive affect	Intercept	.646	.021	2218.5	30.78	p<.001
	Pos	.170	.011	41071.4	15.27	p<.001
	Sex=Male	-.083	.021	2321.7	-3.99	p<.001
	Engfl	.057	.009	2168.8	5.98	p<.001
	Sex=Male * Pos	.058	.019	41071.1	3.04	.002
Interpersonal Problems	Intercept	.183	.025	1368.9	7.23	p<.001
	Int	-.224	.012	11917.7	-19.10	p<.001
	Sex=Male	-.219	.026	1358.4	-8.57	p<.001
	Engfl	.028	.011	1350.2	2.49	.013
	Sex=Male * Int	.109	.020	11903.3	5.43	p<.001

Error rate set to .001

Table D.24: Semantics I – 5 Predictor Model

Feature	Parameter	Estimate	SE	df	t	p
Interpersonal Problems	Intercept	.145	.026	1336.98	5.54	p<.001
	Int	-.129	.020	12033.27	-6.32	p<.001
	Sex=Male	-.216	.026	1338.19	-8.45	p<.001
	Engfl	.050	.012	1337.13	4.22	p<.001
	Sex=Male * Int	.107	.020	12007.74	5.37	p<.001
	Engfl*Int	-.052	.009	12020.75	-5.63	p<.001

Error rate set to .001

Table D.25: Semantics II – 1 Predictor Models

Feature	Parameter	Estimate	SE	df	t	p
Polysemy-Word net, manual	Intercept	-.114	.014	3993.6	-8.15	p<.001
	Pol_wn	.004	.001	7851.9	4.28	p<.001
Polysemy-framenet, manual	Intercept	.065	.014	2841.4	4.64	p<.001
	Pol_fn	-.024	.002	16663.5	-13.11	p<.001
Word frequency-all	Intercept	-.159	.021	6170.3	-7.56	p<.001
	Frq	.027	.006	18276.7	4.35	p<.001
Word familiarity-content	Intercept	-.881	.121	12687.1	-7.30	p<.001
	fam	.002	.000	13884.3	7.96	p<.001
Concreteness-content	Intercept	.499	.033	11993.2	14.94	p<.001
	Cnc	-.001	.000	14499.5	-15.99	p<.001
Imageability-content	Intercept	.456	.035	9218.6	13.17	p<.001
	IMG	-.001	.000	17465.5	-14.91	p<.001
Meaningfulness	Intercept	-.177	.026	4782.2	-6.84	p<.001
	Mea	.000	.000	14653.5	4.41	p<.001
Polysemy-Coh-Metrix	Intercept	-.057	.016	4799.1	-3.70	p<.001
	Pol_c	-.004	.002	9178.6	-2.24	.025
Hypernymy-nouns	Intercept	.102	.013	2739.6	7.98	p<.001
	Hyp_n	-.019	.001	26021.5	-18.14	p<.001
Hypernymy-Verbs	Intercept	-.090	.013	3217.4	-6.94	p<.001
	Hyp_v	.059	.004	4522.3	13.36	p<.001

Error rate set to .001

Table D.26: Semantics II – 3-Predictor Models

Feature	Parameter	Estimate	SE	df	t	p
Polysemy-Wordnet	Intercept	.522	.013	17805.9	38.88	p<.001
	Pol_wn	.010	.001	6351.1	10.86	p<.001
	Sex=Male	-.079	.011	11182.3	-7.52	p<.001
	Engfl	.051	.005	11180.7	10.35	p<.001
Polysemy- framenet	Intercept	.745	.022	2669.5	33.89	p<.001
	Pol_fn	-.022	.002	11819.4	-10.02	p<.001
	Sex=Male	-.069	.020	2166.5	-3.38	.001
	Engfl	.058	.009	2166.4	6.08	p<.001
Frequency	Intercept	.442	.023	14306.5	19.53	p<.001
	Frq	.059	.007	12169.1	8.34	p<.001
	Sex=Male	-.081	.011	11218.0	-7.66	p<.001
	Engfl	.051	.005	11216.1	10.49	p<.001
Familiarity	Intercept	-1.070	.144	11650.9	-7.42	p<.001
	fam	.003	.000	11697.5	11.64	p<.001
	Sex=Male	-.083	.011	11147.3	-7.85	p<.001
	Engfl	.052	.005	11145.7	10.62	p<.001
Concreteness	Intercept	1.967	.040	12000.2	49.64	p<.001
	Cnc	-.004	.000	11433.5	-38.23	p<.001
	Sex=Male	-.069	.020	2166.6	-3.39	.001
	Engfl	.057	.009	2166.5	6.08	p<.001
Imageability	Intercept	.483	.041	8119.2	11.68	p<.001
	WRD_IMG_c	-.001	.000	17705.2	-15.07	p<.001
	Sex=Male	-.191	.025	2323.1	-7.49	p<.001
	Engfl	.016	.012	2323.1	1.33	.182
Meaningfulness	Intercept	.660	.028	14100.4	23.92	p<.001
	Mea	.0002	.000	12412.9	-2.22	.027
	Sex=Male	-.079	.011	11205.9	-7.47	p<.001
	Engfl	.052	.005	11204.3	10.54	p<.001
Hypernymy, Nouns	Intercept	.661	.021	2405.7	30.87	p<.001
	Hyp_n	.004	.001	13998.5	3.13	.002
	Sex=Male	-.069	.020	2166.3	-3.38	.001
	Engfl	.057	.009	2166.2	6.05	p<.001
Hypernymy, verbs	Intercept	-.066	.026	2575.4	-2.53	.012
	Hyp_v	.058	.004	4531.4	13.21	p<.001
	Sex=Male	-.172	.025	2360.0	-6.86	p<.001
	Engfl	.019	.012	2359.9	1.60	.109

Error rate set to .001

Table D.27: Semantics II – English Fluency by Feature Interaction Model

Feature	Parameter	Estimate	SE	df	t	p
Polysemy-wordnet	Intercept	.521	.019	17178.8	26.85	p<.001
	Pol_wn	.010	.002	6315.3	5.40	p<.001
	Sex=Male	-.079	.011	11182.3	-7.52	p<.001
	Engfl	.051	.009	16731.2	5.56	p<.001
	Engfl * Pol_wn	.000	.001	6321.2	-.05	.960
Polysemy-framenet	Intercept	.729	.025	4397.0	28.93	p<.001
	Pol_fn	-.017	.005	11817.0	-3.80	p<.001
	Sex=Male	-.069	.020	2166.5	-3.38	.001
	Engfl	.067	.012	4707.7	5.71	p<.001
	Engfl * Pol_fn	-.003	.002	11818.4	-1.32	.187
Frequency	Intercept	.498	.042	10726.6	11.92	p<.001
	FRQ	.039	.014	12133.0	2.74	.006
	Sex=Male	-.081	.011	11217.4	-7.66	p<.001
	Engfl	.021	.020	10576.3	1.03	.303
	Engfl * FRQ	.011	.007	12128.4	1.57	.116
Familiarity	Intercept	-.746	.291	11577.6	-2.56	.010
	Fam	.002	.000	11703.2	4.64	p<.001
	Sex=Male	-.083	.011	11147.2	-7.86	p<.001
	Engfl	-.133	.141	11574.5	-.95	.344
	Engfl * Fam	.000	.000	11704.0	1.32	.188
Concreteness	Intercept	1.190	.065	10146.3	18.26	p<.001
	Cnc	-.002	.000	7817.8	-9.05	p<.001
	Sex=Male	-.076	.010	11282.8	-7.28	p<.001
	Engfl	.218	.031	10100.7	6.94	p<.001
	Engfl * Cnc	-.001	.000	7822.6	-5.40	p<.001
Hypernymy for nouns	Intercept	.638	.015	14331.8	43.40	p<.001
	Hyp_n	-.013	.002	24910.0	-5.06	p<.001
	Sex=Male	-.079	.011	11153.4	-7.52	p<.001
	Engfl	.071	.007	13813.5	10.35	p<.001
	Engfl * Hyp_n	-.005	.001	24912.8	-4.08	p<.001

Error rate set to .001

Table D.28: Semantics II – Sex by Feature Interaction Model

Feature	Parameter	Estimate	SE	df	t	p
Polysemy-WN	Intercept	.550	.015	18756.1	37.68	p<.001
	Pol_wn	.007	.001	6345.0	5.93	p<.001
	Sex=Male	-.161	.020	16789.8	-8.12	p<.001
	Engfl	.051	.005	11179.1	10.36	p<.001
	Sex=Male * Pol_wn	.010	.002	6360.0	4.85	p<.001
Polysemy-FN	Intercept	.711	.023	3028.3	31.32	p<.001
	Pol_fn	-.010	.003	41070.4	-3.45	.001
	Sex=Male	.025	.025	5175.4	.97	.335
	Engfl	.057	.009	2168.8	5.98	p<.001
	Sex=Male * Pol_fn	-.030	.005	41070.3	-6.30	p<.001
Frequency	Intercept	.580	.027	12507.7	21.76	p<.001
	FRQ	.010	.009	11892.2	1.14	.254
	Sex=Male	-.479	.043	10342.6	-11.16	p<.001
	Engfl	.051	.005	11199.4	10.45	p<.001
	Sex=Male * FRQ	.141	.015	11884.0	9.55	p<.001
Familiarity	Intercept	-.214	.177	11513.5	-1.21	.226
	Fam	.001	.000	11607.4	4.63	p<.001
	Sex=Male	-2.729	.302	11469.6	-9.04	p<.001
	Engfl	.052	.005	11148.0	10.62	p<.001
	Sex=Male * Fam	.448	.051	11610.1	8.77	p<.001
Concreteness	Intercept	1.542	.041	10897.6	37.96	p<.001
	Cnc	-.003	.000	7818.4	-23.81	p<.001
	Sex=Male	-.223	.068	10083.9	-3.29	.001
	Engfl	.050	.005	11279.7	10.36	p<.001
	Sex=Male * Cnc	.000	.000	7811.1	2.20	.028
Hypernymy, n	Intercept	.680	.012	15293.3	54.94	p<.001
	Hyp_n	-.023	.002	24887.7	-15.11	p<.001
	Sex=Male	-.097	.015	13823.4	-6.53	p<.001
	Engfl	.051	.005	11145.4	10.53	p<.001
	Sex=Male * Hyp_n	.004	.003	24863.7	1.67	.096

Error rate set to .001

Table D.29: GLMM-1 Predictor

Feature	Model Term	Estimate	SE	t	p	Exp (Estimate)
Avg. Word Length	Intercept	.324	.026	12.37	p<.001	1.383
	WL_ltm	-.154	.006	-26.14	p<.001	.857
Prepositions	Intercept	-.337	.014	-24.08	p<.001	.714
	Prep	1.226	.061	20.18	p<.001	3.409
Relative clauses	Intercept	-.377	.015	-25.71	p<.001	.686
	Rel	.532	.013	42.66	p<.001	1.702
Cultural Loading	Intercept	-.143	.013	-11.14	p<.001	.867
	Cult	-.326	.010	-34.19	p<.001	.721

Error rate set to .001
Avg. Word Length: Average number of letters in each word in the item

Table D.30: GLMM – 3 Predictor

Feature	Model Term	Estimate	SE	t	p	Exp (Estimate)
Avg. Word Length	Intercept	.221	.037	6.04	p<.001	1.248
	WL_ltm	-.155	.006	-26.11	p<.001	.856
	Sex=Male	-.097	.027	-3.62	p<.001	.907
	Engfl	.075	.012	6.22	p<.001	1.078
Prepositions	Intercept	-.438	.029	-15.13	p<.001	.646
	Prep	1.23	.061	20.12	p<.001	3.421
	Sex=Male	-.09	.026	-3.45	.001	.914
	Engfl	.071	.012	6.01	p<.001	1.074
Relative Clauses	Intercept	-.456	.027	-17.09	p<.001	.634
	Rel	.536	.013	42.82	p<.001	1.709
	Sex=Male	-.077	.023	-3.31	.001	.926
	Engfl	.056	.010	5.48	p<.001	1.058
Cultural Loading	Intercept	-.219	.028	-7.70	p<.001	.804
	SEM_Cult	-.327	.010	-34.04	p<.001	.721
	Sex=Male	-.085	.027	-3.20	.001	.918
	Engfl	.057	.012	4.70	p<.001	1.059

Error rate set to .001
Avg. Word Length: Average number of letters in each word in the item

Table D.31: GLMM – English fluency by feature interaction

Feature	Model Term	Estimate	SE	t	p	Exp (Estimate)
Avg. Word Length	Intercept	.146	.055	2.68	.007	1.158
	WL_ltm	-.134	.012	-11.03	p<.001	.874
	Sex=Male	-.097	.027	-3.61	p<.001	.907
	WL_ltm*Engfl	-.011	.006	-1.95	.051	.989
	Engfl	.115	.024	4.70	p<.001	1.122
Prepositions	Intercept	-.446	.031	-14.39	p<.001	.64
	Prep	1.343	.125	10.79	p<.001	3.832
	Sex=Male	-.09	.026	-3.45	.001	.914
	Prep*Engfl	-.06	.058	-1.05	.294	.941
	Engfl	.075	.013	5.76	p<.001	1.078
Relative Clause	Intercept	-.512	.032	-15.94	p<.001	.6
	Rel	.646	.026	25.22	p<.001	1.907
	Sex=Male	-.076	.023	-3.25	.001	.927
	Engfl*Rel	-.059	.013	-4.58	p<.001	.943
	Engfl	.086	.014	6.15	p<.001	1.09
Cultural Loading	Intercept	-.202	.029	-7.10	p<.001	.817
	Cult	-.45	.019	-23.17	p<.001	.638
	Sex=Male	-.085	.027	-3.21	.001	.918
	Cult*Engfl	.065	.009	7.26	p<.001	1.067
	Engfl	.048	.012	3.90	p<.001	1.05

Error rate set to .001
Avg. Word Length: Average number of letters in each word in the item

Table D.32: GLMM– Sex by Feature Interaction

Feature	Model Term	Estimate	SE	t	p	Exp (Estimate)
Avg. Word Length	Intercept	.228	.040	5.75	p<.001	1.256
	WL_ltm	-.157	.007	-22.16	p<.001	.855
	Sex=Male	-.119	.057	-2.08	.037	.888
	WL_ltm*Sex=Male	.006	.013	.46	.642	1.006
	Engfl	.075	.012	6.21	p<.001	1.078
Prepositions	Intercept	-.431	.029	-14.76	p<.001	.65
	Prep	1.133	.073	15.44	p<.001	3.104
	Sex=Male	-.111	.029	-3.82	p<.001	.895
	Prep*Sex=Male	.3	.132	2.27	.023	1.35
	Engfl	.071	.012	6.00	p<.001	1.074
Relative clause	Intercept	-.444	.027	-16.19	p<.001	.641
	Rel	.514	.015	33.66	p<.001	1.672
	Sex=Male	-.111	.031	-3.63	p<.001	.895
	Rel*Sex=Male	.067	.027	2.53	.011	1.07
	Engfl	.056	.010	5.45	p<.001	1.058
Cultural Loading	Intercept	-.221	.029	-7.75	p<.001	.802
	Cult	-.313	.012	-27.22	p<.001	.731
	Sex=Male	-.079	.027	-2.92	.004	.924
	Cult*Sex=Male	-.045	.021	-2.18	.029	.956
	Engfl	.057	.012	4.70	p<.001	1.059

Error rate set to .001
Avg. Word Length: Average number of letters in each word in the item

Table D.33: GLMM–5 Predictor

Feature	Model Term	Estimate	SE	t	p	Exp (Estimate)
Avg Word length	Intercept	.153	.057	2.70	.007	1.166
	WL_ltm	-.136	.013	-10.67	p<.001	.873
	Sex=Male	-.119	.057	-2.08	.037	.888
	Engfl	.115	.024	4.70	p<.001	1.122
	WL_ltm*Sex=Male	.006	.013	.470	.639	1.006
	WL_ltm*Engfl	-.011	.006	-1.95	.051	.989
Preposition	Intercept	-.439	.031	-14.07	p<.001	.645
	Prep	1.246	.129	9.68	p<.001	3.477
	Sex=Male	-.111	.029	-3.82	p<.001	.895
	Engfl	.075	.013	5.76	p<.001	1.078
	Prep*Sex=Male	.3	.132	2.28	.023	1.35
	Prep*Engfl	-.06	.057	-1.05	.292	.941
Relative clause	Intercept	-.5	.033	-15.29	p<.001	.607
	Rel	.624	.027	23.38	p<.001	1.866
	Sex=Male	-.11	.031	-3.59	p<.001	.896
	Engfl	.086	.014	6.12	p<.001	1.089
	Rel*Sex=Male	.067	.026	2.55	.011	1.069
	Rel*Engfl	-.059	.013	-4.58	p<.001	.943
Cultural Loading	Intercept	-.204	.029	-7.15	p<.001	.815
	Cult	-.435	.020	-21.38	p<.001	.647
	Sex=Male	-.079	.027	-2.93	.003	.924
	Engfl	.048	.012	3.90	p<.001	1.05
	Cult*Sex=Male	-.046	.021	-2.20	.028	.956
	Cult*Engfl	.065	.009	7.28	p<.001	1.067

Error rate set to .001

Avg. Word Length: Average number of letters in each word in the item

D.5 Multiple Feature Models

Table D.34: Revised Full Model

Domain	Parameter	Estimate	Std. Error	df	t	p
-	Intercept	.496	.023	16737.1	21.49	p<.001
Length	Word Length	.101	.007	17154.2	15.12	p<.001
Morphology I	Nouns	-.770	.079	14443.3	-9.71	p<.001
Morphology II	Bound Morphemes	-.165	.007	14204.9	-23.53	p<.001
Syntax	Relative Clauses	.560	.012	7714.1	45.80	p<.001
Semantics I	Cultural loading	-.237	.008	22975.8	-29.74	p<.001
Semantics I	Positive Affect	.155	.011	18061.5	14.62	p<.001
Semantics II	Hypernymy, Nouns	.013	.003	12152.5	4.35	p<.001

Table D.35: Revised Full Model with Sex and English fluency

Domain	Parameter	Estimate	Std. Error	df	t	p
-	Intercept	.451	.029	9974.2	15.70	p<.001
Length	Word Length	.101	.007	17053.4	15.19	p<.001
Morphology I	Nouns	-.776	.079	14344.2	-9.79	p<.001
Morphology II	Bound Morphemes	-.165	.007	14157.3	-23.57	p<.001
Syntax	Relative Clauses	.559	.012	7714.8	45.68	p<.001
Semantics I	Cultural loading	-.237	.008	22828.5	-29.70	p<.001
Semantics I	Positive Affect	.156	.011	18077.8	14.71	p<.001
Semantics II	Hypernymy, Nouns	.013	.003	12141.4	4.45	p<.001
-	Sex=Male	-.109	.019	2602.0	-5.76	p<.001
-	EngFl	.045	.009	2601.9	5.12	p<.001

Appendix E

Figures

Figure E.1: Histogram of CES-D Composite Scores

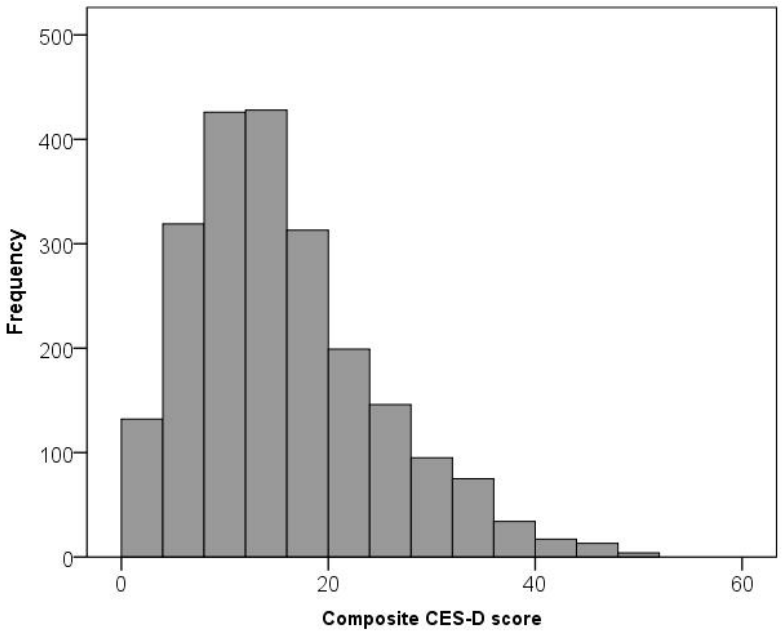


Figure E.2: Histogram of CES-D Composite Scores, by Level of English Fluency

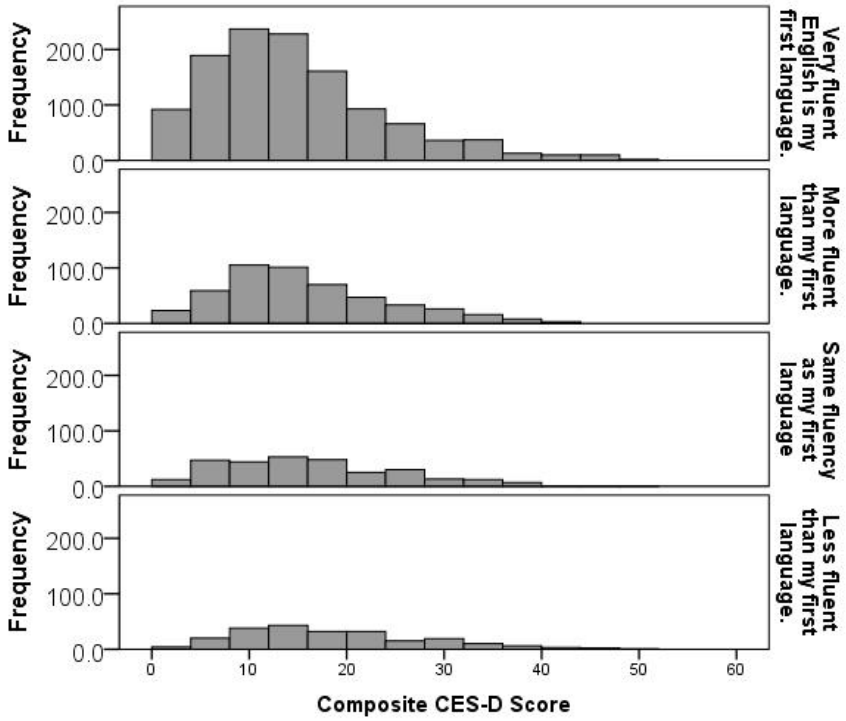
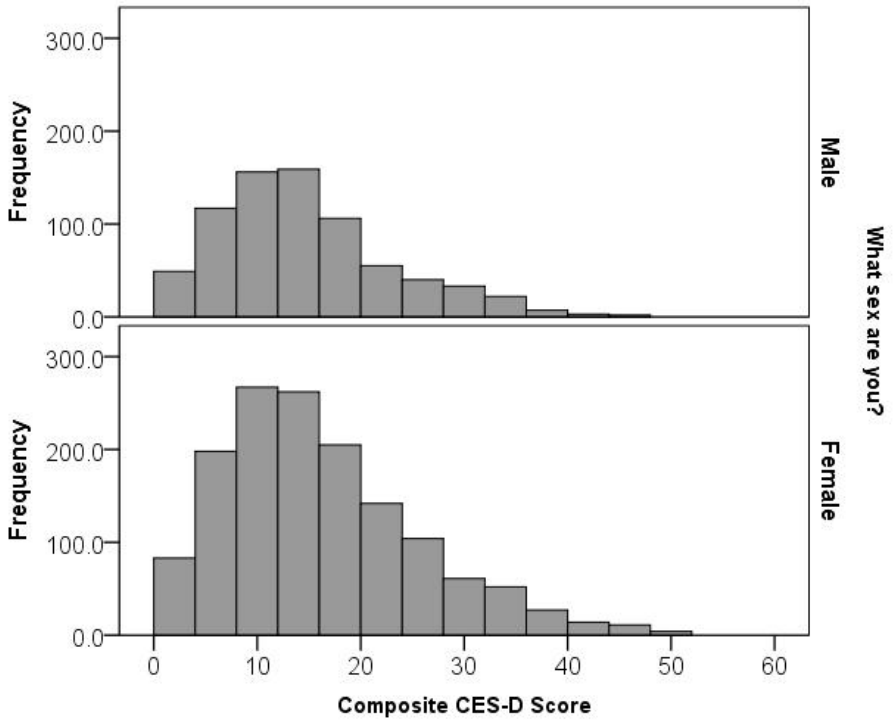


Figure E.3: Histogram of CES-D Composite Scores, by Sex



Appendix F

Syntax Examples

F.1 Linear Mixed Model Syntax

1-Predictor Model, Word Count

```
MIXED CESD BY sex_rpt WITH engfl_rpt DES_WC_nw DES_WL_ltm SYN_REL SYN_Poss
 /CRITERIA=CIN(95) MXITER(15) MXSTEP(10) SCORING(1) SINGULAR(0.000000000001)
 HCONVERGE(0, ABSOLUTE) LCONVERGE(0, ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
 /FIXED=DES_WC_nw | SSTYPE(3)
 /METHOD=REML
 /PRInt=CORB COVB G HISTORY(1) LMATRIX R SOLUTION TESTCOV
 /REPEATED=Index | SUBJECT(new_casenum) COVTYPE(TPH).
```

F.2 Generalized Linear Mixed Model Syntax

3-Predictor Model, Word Length, letters

```
GENLINMIXED
 /DATA_STRUCTURE SUBJECTS=new_casenum REPEATED_MEASURES=Index
 COVARIANCE_TYPE=COMPOUND_SYMMETRY
 /FIELDS TARGET=CESD TRIALS=NONE OFFSET=NONE
 /TARGET_OPTIONS DISTRIBUTION=POISSON LINK=LOG
 /FIXED EFFECTS=DES_WL_ltm  sex_rpt engfl_rpt USE_IntERCEPT=TRUE
 /BUILD_OPTIONS TARGET_CATEGORY_ORDER=ASCENDING INPUTS_CATEGORY_ORDER
 =ASCENDING MAX_ITERATIONS=100 CONFIDENCE_LEVEL=95 DF_METHOD=RESIDUAL
 COVB=ROBUST PCONVERGE=0.000001(ABSOLUTE) SCORING=0
 SINGULAR=0.000000000001.
```