

Meaningful Nonsense: Invented Words Reveal Characteristics of Emotional
Stimuli

Emil G. Moldovan

Columbus, OH

Bachelor of Science. Ohio State University, 2011

A Thesis presented to the Graduate Faculty
of the College of William and Mary in Candidacy for the Degree of
Master of Arts

Department of Psychology

The College of William and Mary
August, 2014

APPROVAL PAGE

This Thesis is submitted in partial fulfillment of
the requirements for the degree of

Master of Arts



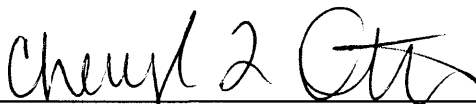
Emil Moldovan

Approved by the Committee, July, 2014

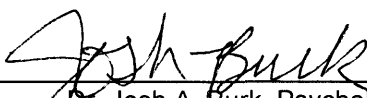


Committee Chair

Dr. Todd M. Thrash, Psychology
The College of William and Mary



Dr. Cheryl L. Dickter, Psychology
The College of William and Mary



Dr. Josh A. Burk, Psychology
The College of William and Mary

COMPLIANCE PAGE

Research approved by

Protection of Human Subjects Committee

Protocol number(s): PHSC-2012-11-11-8274-tmthra

PHSC-2014-01-27-9256-tmthra

Date(s) of approval: 11-11-2012

01-27-2014

ABSTRACT

Traditional linguistic theories hold that the meaning of words is totally culturally determined. Sound-symbolism research, on the other hand, suggests that particular sounds have intrinsic meaning. We assessed this theory by asking participants to invent words to describe stimuli with known emotional features. Coders then assessed these invented words for valence and arousal, two orthogonal dimensions of emotional meaning. Multilevel structural equation modeling was used to show that picture features predicted the features of invented words. These effects were present after dropping two different sorts of invalid entries. The effects also subsisted after we accounted for the degree to which invented words resemble English words with emotional meaning. The results have implications for theories of linguistic relativism, scientists interested in measuring implicit affect, and individuals who experience emotions for which they lack words.

TABLE OF CONTENTS

Acknowledgements	ii
List of Figures	iii
Chapter 1. Introduction	1
Chapter 2. Methods: Main Study	7
Chapter 3. Methods: Validity Study	15
Chapter 4. Results: Validity Study	31
Chapter 5. Results: Main Study	33
Chapter 6. Confound Analysis	38
Chapter 7. Discussion	43
References	53

ACKNOWLEDGEMENTS

I wish to express my appreciation to Professor Todd Thrash, under whose guidance this investigation was conducted, for his patience, guidance and criticism throughout the investigation. I am also indebted to Professors Cheryl Dickter and Josh Burke for their careful reading and criticism of the manuscript. Further, I am grateful to David Newman, a fellow graduate student, for his help writing SPSS code as part of the validity study included here. Prof. Thrash's lab manager, Amanda Fuller, helped prepare some of the experimental material used in the main study. Michael LeFew helped me with an important literature review for the main study. Finally, I am grateful to all of the research assistants who coded invented words. A tedious task that was, but an extremely useful one!

LIST OF FIGURES

1. Cross-Classified Structural Equation Model for Main Study	50
2. Measurement Model to Construct the Dictionary Used in the Confound Analysis	51
3. Cross-Classified Structural Equation Model for the Confound Analysis	52

Introduction

The theory of core affect holds that two basic variables, valence and arousal, define the most relevant dimensions of the abstract space within which emotional experiences operate (Yik, Russell & Steiger, 2011). For any given language, linguistic symbols for emotional experiences only represent certain regions of this two-dimensional space. Different cultures have assemblages of linguistic tokens that pick out different regions in this two-dimensional space (Russell, Lewicka & Niit, 1989). The reason for this, it is assumed, is that the relationship between language and emotional experience is culturally-determined (Barrett, Lindquist & Gendron, 2007). This means that the meaning associated with any one linguistic token is arbitrary.

Previous theorists have focused mostly on utterances with meaning that is culturally derived (e.g., words like "happy," "calm"). We introduce invented words as a new type of utterance. Invented emotional words are words that people invent to designate an experience for which they may or may not have culturally derived words to represent. In this thesis, I aim to demonstrate empirically that invented words operate in very similar ways as culturally-derived emotion words do. That is, I aim to show that the same two-dimensional structure explains the relationship between invented emotional words and their referents as the two-dimensional structure that explains the relationship between natural language words and their referents. Two types of meaning will be considered: meaning that any one person ascribes to the words they produce, and meaning that people ascribe to words produced by others. The methodology I will employ has the

ability demonstrate that these new types of utterances have both of these types of meaning.

The Theory of Core Affect

The theory of core affect is often explained in juxtaposition to theories of basic emotions (Ekman, 1992). Theories of basic emotions posit that there are two types of emotions: basic, and complex. Each basic emotion is qualitatively different from every other basic emotion in several ways. Each basic emotion feels different, has a unique neurological underpinning, differs in facial expression, impacts cognition in unique ways, and plays a unique evolutionary function. While basic emotions are atomic in that they cannot be broken down into more basic emotions, complex emotions are blends of basic emotions.

The theory of core affect is different from this in two ways. According to this view, rather than positing that different emotional states are qualitatively different, a few broad dimensions underlie all emotional experiences. These underlying dimensions are referred to as "core affect." Differences in emotions are thus quantitative, not qualitative. The two main dimensions are valence and arousal. Valence refers to the hedonic value of an emotional experience; hedonic value ranges from "positive" and "negative." Arousal refers to the energy level of the emotional experience, and varies from "calm" to "aroused." Core affect theorists claim that these dimensions explain many of the facets of emotions including their qualia, their neurological instantiation, and their facial expression.

According to the theory of core affect, different cultures will have different conceptions about which emotions are "basic." Some cultures will not even posit

any emotions as basic. Differences in conception of emotions are so dramatic that some cultures might not even have words for emotions that other cultures construe to be basic. For example, according to Wierzbicka (2009), Russian does not have a word that stands for the English near-equivalent of "surprise." In English, "surprise" occurs upon noticing something novel that can be pleasant, or unpleasant. But in Russian, the closest term to "surprise" denotes novel unpleasant occurrences; there is no term that refers to novelty in a way that is devoid of valence connotations.

Despite cultural differences in which emotion words are available to speakers, core affect researchers have demonstrated that the same two-dimension structure underlies the emotional lexicon in many unrelated languages (Russell et al., 1989). Valence and arousal delineate the conceptual possibilities that can be represented in language, not what has actually been represented in any one language. Even though Russian speakers do not have a word for a surprised state that may be either positive or negative, they can still experience that state, and they can conceptualize it (albeit not using single Russian words).

While the general structure of emotional experience and language use is universal, the specific words used to designate particular regions of the valence/arousal space may not be. According to a theory that I will refer to as linguistic relativism (e.g., views espoused in Hockett, 1960), there is nothing intrinsic to any given word that will make it well-suited to describe a given experience. The meaning of all emotion words is totally culturally ascribed. (There are many other theories commonly referred to as "linguistic relativism")

that will not be discussed in this document).

If the meaning of words is culturally ascribed this implies that if a person lacks any cultural training in the meaning of a word, then he or she is incapable of accurately ascribing meaning to it. Emotion words from a language that is unrelated to one's own look senseless - they have no meaning. Also, if a person invents an emotion word without telling anyone what that emotion word means, then nobody should be able to accurately decode the meaning of that word. S/he will not be able to tell, based on that word alone, what state another person was in when s/he produced it, or what state that word refers to. Both of these hypotheses straightforwardly flow out of the linguistically relativistic view. But another group of theorists would make a different prediction.

Sound Symbolism

The sound-symbolism hypothesis is that some forms of meaning are carried forth by sub-word units (Nuckolls, 1999). Sound symbols can be single phonemes, or combinations of phonemes. A phoneme is simply a sound that a person produces as he or she pronounces a word. In either case, a sound symbol is not a stand-alone word. An example of a sound-symbol is the sound that the letter "y" produced when an American English speaker utters the word "happy."

Sound symbolism researchers have found sounds that covary strongly with a variety of types of meaning. Some connote natural phenomena (e.g., words with the sound-symbol "gl" often connote light), some connote body parts (e.g., words with the sound-symbols "sn" often convey ideas related to the nose),

and some connote emotion (e.g., the sound "ee" is often perceived as connoting happy states).

When a listener hears a sound-symbol embedded in a word, the referent of the sound-symbol comes to his or her mind along with the meaning of the word as a whole. If meaning resides at the sub-word level, then combining these sub-word units in novel ways would still allow their meaning to be conveyed. So a person might be able to hear a totally novel word and still be able to decode its meaning if s/he can intuit the meaning of its sound symbols.

There are two versions of the sound-symbolism hypothesis (Weiss, 1964). One of them is consistent with linguistic relativism, while the other is not. The "soft" formulation holds that the meaning of a sound-symbol can be culturally learned. An example of such a sound symbol is the sound "gl;" it is unlikely that the sound "gl" is associated with light in any other language than English. Many current researchers have replicated and extended versions of Weiss' original work. For example, Parault and Schwanenflugel (2006) showed that English speakers better learn English words when their phonetic qualities are sound-symbolically congruent with their semantic meaning.

The "hard" formulation of the sound-symbolism hypothesis holds that the meaning of a sound-symbol is culturally independent. This formulation is not consistent with linguistic relativism. It holds that some words are better matches for some emotional states than other words are. Supposedly, humans are born capable of understanding the meaning of certain sounds without any cultural training. In support of this formulation, Koriat, (1975) showed that individuals can

guess with above-chance accuracy the meaning of emotion words in language that they do not speak. More current researchers, such as Myers-Schulz, Pujara, Wolf and Koenigs (2013), produced nonwords with sound symbols embedded in them which are hypothesized to have a particular connotation, and show that participants indeed perceive certain sounds to be systematically linked to certain emotional connotations.

Main study

Both formulations of the sound symbolism hypothesis hold that subword units carry meaning. Thus, both formulations would predict that listeners can pick up on the emotional meaning of invented words that carry the sound-symbols that are found in other emotion words of their language. There are two further implications in light of the dimensional theory of affect. First, it should be possible to organize the meaning of emotional sound-symbols into valence/arousal space. Second, it should also be possible to organize invented words that have been derived from these sound-symbols in this space. If this turns out to be the case, it will not support either formulation of the sound-symbolism hypothesis more than the other.

To our knowledge, no study into the emotional role of sound-symbols has explored their function in word production. Most sound-symbolism researchers have not considered them in light of modern theories of affect. Further, most studies employ either preexisting words from a language that is unknown to the participant or experimenter-invented nonwords.

To test whether humans can code emotional information into invented

words, I asked participants to invent words to describe pictures with known emotional characteristics. Then, I asked a group of coders to rate the invented words along the valence and arousal dimensions. Given that valence and arousal are orthogonal dimensions, I expect that picture valence will predict nonword valence independent of picture arousal. Likewise, I expect that picture arousal will predict nonword arousal independent of picture valence. To note, the results of this study will be generalizable along two independent dimensions. The random sampling of participants, along with the utilization of a statistical model that includes random effects for participants, will make the results generalizable to other participants. The random sampling of stimuli, along with the utilization of a statistical model that also includes random effects for stimuli, will make the results generalizable to other stimuli (Judd, Westfall, & Kenny, 2012).

Methods for Main Study

Nonword Generation

Participants

55 William and Mary undergraduate students completed the study. All participants were at least 18 (mean age: 18.85). 44 were Caucasian and 42 were male.

The university Institutional Review Board ensured that the study is ethical and fair to participants. Students were instructed not to participate in the study if they were adverse to viewing graphic depictions of sex or violence. Including such pictures was necessary and warranted. It was necessary to include graphic pictures because pictorial stimuli are not ideal for the elicitation of strong

emotions. I needed to elicit strong emotions in order to ensure that the study is externally valid. The inclusion was warranted due to its low risk. The more shocking pictures from the stimulus set were used in previous studies, with no notable repercussions. Still, to mitigate even the lowest levels of distress, I included two more highly positive pictures after the end of the study. This procedure helps ensure that any negative affect experienced during the experiment is not carried by the participants beyond the experimental session. The data produced for these last two pictures were not included in any of the analyses.

Material

I selected 66 Pictures from the International Affective Picture System (IAPS; Lang, Bradley & Cuthbert, 1999). The IAPS consists of thousands of pictures designed to elicit a variety of emotional reactions. To measure these reactions, Lang et al. (1999) asked participants to rate each picture along the valence and arousal dimensions using the SAM method. SAM is a visual representation of valence (ranging from a frowning to a smiling cartoon figure) and arousal (ranging from a serene to an agitated cartoon figure) that participants use to indicate their emotional reactions to any one picture.

I sampled this stimulus set in a way that maximized both the external validity of our stimulus subset and the statistical power to pick up the effects of interest. I considered the emotions elicited by the pictures, the content of the pictures, and other picture characteristics.

Picture Affect. The 66 pictures cover the full range of affective experiences

that the IAPS set as a whole does. Plotted in the valence/arousal space, they produce the same "v" shape as IAPS set. The range of emotion encompasses both extreme emotion (i.e., high levels of activated negative affect and high levels of activated positive affect) as well as neutral emotions (low arousal). This selection procedure helps ensure that our results are valid for a wide array of emotional states. Although stimuli were sampled randomly, the affective space was not sampled randomly. First, I selected pictures to maximize the variance for both the valence and arousal dimensions of the stimulus set as a whole. The primary goal behind this decision was to maximize the statistical ability to pick up on effects of interest. Second, with the exception of erotic nudes, I chose pictures with similar norms from men and women. Third, I only chose pictures with univalent emotional connotations. Thus, for example, I eschewed pictures that displayed sexualized violence due to the possibility that some participants would react more strongly to the negative content of the picture, while others might be aroused by the sexual elements of the picture. Selecting for more homogenous response styles might lead to a reduction in the error variance of the resulting data.

Picture Content. Three pictures were selected to represent each of the 18 different topical categories introduced in (Bradley, Codispoti, Cuthbert & Lang, 2001). Eight of these categories are pleasant (nature, families, food, sports, adventure, attractive men, attractive women, erotic couples), two are neutral (household objects, mushrooms), and eight are unpleasant (pollution, illness, loss, accidents, contamination, attacking animals, attacking humans, mutilated

bodies). In addition, I added six more neutral household objects picture in order to satisfy criteria discussed in "picture emotions," above.

For those categories that involved humans, two pictures displayed males and one displayed females. An exception to this is attractive men and attractive women, for which I selected three male and three female erotic nudes. Further, those categories that involve humans are also racially diverse, in that for any one category two pictures display white individuals and one picture displays black individuals.

Picture Characteristics. I rated pictures for luminosity, complexity, whether focal elements of the picture are in the foreground or in the background, and whether the picture displayed humans or animals. Stimuli were chosen to ensure that items with positive and items with negative emotional connotations matched across all these dimensions. These dimensions were not central to the project, so ratings for them were done intuitively by the researchers alone.

Experimental Procedure

Participants first completed a series of questionnaires about state and trait emotional experiences, mindfulness, and impression management. As none of these questionnaires will be integrated into the analysis of this paper, they will not be further discussed.

Participants were instructed that they were to invent a nonword that seems fitting to each picture. The instructions to create nonwords were thus:

... we are interested in how people invent new words, or “nonsense words.” Here are a few examples (but please don't use these): squimf,

alfrastic, zscromed, div, seffle.

... For each picture, your task is to make up a new word that seems fitting to that picture—that is, come up with a word that seems to capture the essence of what is depicted in the picture. It is important that you not base your words on existing words from English or any other language. The words should also not look like any English word you know or any other word you know. For example, "computery" looks too much like "computer" and would thus be an inappropriate response. Instead, these words should be completely new “nonsense words” that seem fitting to the picture based on your subjective impression. So, you are creating your very own, made-up language. No one has ever read or heard this language before. Don't worry about whether the words that you come up with seem unusual. All that matters is that (1) the words are not based on English or any other language and (2) the words seem fitting to the pictures in your view.

...If you're not sure what word would be fitting, go with your gut feeling....

The order of the stimuli was fully randomized. Each picture remained on the screen for six seconds. Participants were instructed to come up with nonwords while the picture is still on the screen. After six seconds, the picture automatically disappeared, giving way to a text entry box into which participants were to enter their nonwords.

Nonword Rating

Some of the coders worked for school credit, while others were paid. In

either case, I selected only coders whose first language was English.

Coders were asked to rate batches of 50 nonwords at a time. These batches were implemented in Qualtrics, an internet-based survey program. Batching words like this was done in order to combat coder fatigue. Although some coders found the task pleasing, many commented on its tedium. So, coders were encouraged to code only as many 50-word batches in one sitting as they felt they could while dedicating their full attention.

Coders used the same SAM scale to code nonwords that the participants from Mikels et al. (2005) used to mark their emotional reactions to IAPS pictures. This facilitated the comparison between direct and indirect measures for emotion. I instructed coders to rate both the valence and arousal connotations for any one nonword before moving on to the next nonword.

Each coder rated every nonword. Instructions for coders closely matched the instructions that the participants in Mikels et al. (2005) were given. I clarified a few additional points. First, I emphasized to coders that they should feel free to use the full range of numerical options at their disposal, including extreme scores. Second, I emphasized to coders the importance of making finely grained distinctions between nonwords of similar, but not identical, affective connotations.

Coders were not told anything about the context within which words were invented. They did not know for which picture any one word was invented, and they did not know that the words were invented to describe pictures. Coders were also not told that the terms they rated were invented to stand for emotional concepts. Indeed, while debriefing at the end of the study it became apparent

that some coders did not even believe that invented words were generated by a human; they thought that the words might have been generated by a machine.

Despite our efforts to stress the importance of this study without actually revealing to coders its true purpose, several coders did not perform their task with due diligence. Specifically, a few coders took too much time to finish their task, failed to convince that they understood the instructions, or otherwise performed the task in a sloppy manner. I decided to drop the data from these coders before examining it. Coding nonwords requires high levels of concentration, and the data of a coder who is not committed is not trustworthy.

Interlude: A Validity Study

As I informally examined the data from the main study, several validity issues became apparent. First, it appears that some participants produced entries by merely pressing keys on a keyboard in a random fashion. Entries such as “sdfby” and “fe;wjafi” are highly unpronounceable and involve characters that are near each other on a standard keyboard. It is possible that entries of this sort were produced by participants who found it too difficult to invent a word for particular pictures. A second validity issue I observed is that despite the clear instructions to participants that they are not to base their invented words on English or any other language, some entries were mere English words, or combinations of English words. Two examples are “californication” and “monkeyingaround.” It is possible that these entries were produced by participants who did not take the study seriously.

The presence of entries that are produced by randomly pressing keys on a keyboard is worrisome. As a participant randomly presses keys on a keyboard, his or her fingers accidentally hit letters in combinations that are highly unlike English. For example, although one might press “zx” together, these two letters never occur together in English. Pronouncing these unusual combinations of letters is unpleasant. Therefore, it is possible that such entries account for a large percentage of the negatively valenced and highly arousing nonwords ratings. Further, it is possible that such entries are predominantly produced as participants view highly negative and arousing pictures.

In order to account for this nuisance mechanism which might drive effects, I need to eliminate entries that are produced by randomly pressing keys on a keyboard. Hereafter, such entries are referred to as “gibberish entries.” However, it is very difficult to reliably differentiate gibberish from valid entries. Many invented words are a little unusual and there is no obvious way to differentiate between an implausible entry and a merely strange one. Before analyzing the results of the main study it is important to derive a gold standard which can be used to objectively determine whether an entry is gibberish or a valid nonword. This gold standard will be established empirically with a validity study.

The first half of the validity study was roughly the same as the main study. The only exception is that I have taken all measures possible for participants to produce valid entries. In the second half of the second study, I asked participants to produce invalid entries by randomly pressing keys. Participants were incentivized to take our instructions to randomly press keys and finish the study

as quickly as possible because the first half of the study was mentally tiring. By comparing the valid with the gibberish entries produced by each participant, I was able to empirically obtain an equation which maximally and validly differentiates between the two.

The presence of entries with English words in the main study is also worrisome. If invented words carry emotional information, this might be due only because invented words contain English words with emotional meaning in them. There are two ways of dealing with this concern. The first approach is to drop entries that are composed entirely of English words. Entries such as “monkeyingaround” straightforwardly carry semantic information. Such entries are not valid data because participants clearly failed to follow the instructions. These entries should therefore be dropped. Entries that resemble English words may or may not be valid. The conservative approach is therefore not to drop them. However, precautions should be taken in order to ensure that, if invented words carry information, it is not due to their resemblance to English words.

Methods of the Validity Study

Participants

Participants were William and Mary undergraduate students. 95 individuals completed the study. Inadvertently, no demographic information was collected for this sample.

Stimuli

The validity study was conducted with a subset of the stimuli that were used in the main study. I selected two pictures from each set of the 18 groups of

pictures discussed above. The stimulus set from the main study reflected the emotional properties of the IAPS as a whole, and each group of pictures inhabited a fairly discrete region of the valence-arousal space. Therefore, this selection procedure ensures that the stimuli in the validity study adequately represent the variation in typical human emotional experience. Both of the experimental conditions in the validity study are comprised of these 44 stimuli. However, due to a technical error, only 42 out of 44 of the stimuli were presented in the valid condition (described below). Each participant saw a different 42 out of the 44 stimuli. Random chance determined which two stimuli any one participant did not experience, so data is missing at random (MAR). All of the statistical procedures used assume that data is MAR.

Procedure

The design of the validity study was similar to that of the main study. In fact, the first half of the validity study was nearly identical to the main study. The instructions were similar in spirit, with two minor changes. First, I framed the task a little bit differently by asking participants to pretend that they are inventing a new language, and that the words they invent should sound fitting not just to them, but to other people as well. The second minor change is that I provided more clear examples of entries that are not acceptable invented terms, and explained what makes these entries unacceptable. Participants had answer three questions about the instructions. They were provided feedback if they answered these questions incorrectly. All participants who continued to the experiment demonstrated that they understood the instructions.

After participants invented words for 42 out of the 44 pictures in the first part of the study, I asked them whether they produced any words that are invalid. Specifically, I explained to participants that it is important for us to know the proportion of terms that they first produced in their mind and then wrote down versus words that they produced by randomly pressing keys on a keyboard. I stressed to participants that they have nothing to lose by admitting that they produced invalid entries because their data is anonymous. I asked the same question again after the second half of the study.

Ideally, participants in the first half of the study admitted to producing entries only by first thinking about them, and participants in the second half of the study admitted to producing entries by randomly pressing keys. In reality, 40% percent of participants in the first half admitted to producing at least some random entries, and 50% percent of participants in the second half admitted to producing at least a few entries which were first thought up and typed in, rather than randomly produced. I dropped the data of participants who admitted to producing more than 20 percent of entries in a way that is invalid relative to each of the conditions. This amounts to 12 participants.

Next, participants read the instructions for the invalid condition. They read that some participants produce invalid data by randomly pressing keys, and that my research group and I, as experimenters, would like their help to understand the attributes of invalid data. In order to do this, I told them, they should quickly progress through the study again and produce invalid data by randomly pressing keys rather than by first inventing a word in their head and then writing it in. I

emphasized that they should feel free to act careless, rushed, or lazy, and to allow any sort of characters to be part of their entries. Participants then answered three questions in order to demonstrate that they understood this change in instructions. Participants who answered these questions incorrectly received corrective feedback before continuing.

Analytic Strategy

The goal of this analysis is to produce a formula which will allow us to predict whether any given case is valid or invalid. This formula will require the use of multilevel cross-classified logistic multilevel modeling. An introduction to this topic is found in Raudenbush and Bryk (2002). Cross-classified multilevel modeling is a form of multilevel regression. Multilevel regression is a form of regression that is used when observations are not independent from each other. Classical regression assumes that the error in predicting results for any one case is independent of the error in predicting the results for any other case. This assumption is often not tenable for situations in which the ways in which observations were collected are dependent upon each other. In this study, any one participant produced many entries. Any given entry that a participant produced is likely to be more similar to other entries that participant produced than it is to entries produced by other participants. Further, different participants produced entries for each picture. Any given entry produced in response to a given picture is likely to be more similar to other entries produced for that same picture than it is to entries produced for other pictures. Multilevel modeling can be used to examine either the average difference between valid and invalid entries

produced for the same picture, or the average difference between valid and invalid entries produced by any one participant. It can also be used to examine how it is that these differences differ for either different stimuli or different participants. To be clear, two different multilevel analyses need to be conducted to take these two different types of nesting into consideration.

Cross-classified multilevel modeling is a type of multilevel modeling that can be used when an observation can be classified in two distinct ways. Each participant invents a nonword for each picture. Therefore, cross-classified multilevel modeling can be used to examine the relationship between picture and nonword characteristics for the average picture and for the average participant (i.e., fixed effects). It can also be used to examine how it is that the relationship between picture and nonword characteristics differs for different pictures, or for different participants (i.e., random effects). Ignoring either level of nesting could lead to systematic biases in standard errors. This in turn can lead to errors in estimating the statistical significance of parameters in the model. Nonsignificant predictors need to be trimmed, so the final model might have the wrong predictors. In a worse-case scenario, a predictor could be left in the model to account for the difference between valid and invalid entries not because it actually differentiates between the two but because variance in this predictor comes from one of the ignored cross-levels of nesting.

Cross-classified multilevel modeling will be combined with binary logistic regression. Binary logistic regression is similar to multiple regression but developed for situations in which the dependent variable is dichotomous.

Independent variables may be continuous or dichotomous. The left side of a logistic regression equation is composed of an intercept and a linear combination of predictor variables. This is similar to multiple regression. The right side of a logistic regression, however, is the logit transformation of the probability that a given case comes from one of two conditions. The logit link is defined as the logarithm of the probability that cases come from one condition divided by 1 minus the probability that cases come from that condition; that is, $\pi/(1-\pi)$. One can take the inverse of the logit in order to determine the probability that any one case comes from one of the conditions. The inverse of the logit function is $1/(1 + e^{-(\text{logit})})$. These transformations allow for the left hand side of the logistic equation (i.e., the linear combination of predictors) to take any real value from negative infinity to positive infinity, and for the probability that any one case comes from any given condition to range from 0 to 1. If the inverse of the logit is greater than 50% for any given case, I predict that it is from the invalid condition.

Multilevel logistic regression is the hierarchical linear modeling extension of logistic regression. This extension allows us to account for the non-independence of results. Specifically, cross-classified logistic multilevel modeling will allow us to account for the multiply nested nature of this study.

Two types of variables will help us distinguish between nonwords and gibberish entries. Some variables take account of the letter composition of any given entry. The letter composition of nonwords and valid entries is likely to be different. Consider "el" and "xg." "El" is more pronounceable than "xg," so entries

that contain “el” may be more likely to be valid, whereas entries that contain “xg” may be more likely to be invalid.

The second type of variable involves the distribution of characters on a standard keyboard. As participants randomly press keys, I hypothesize that some keys are more likely to be pressed than other keys. Characters near the “home” position (i.e., the letters directly underneath one’s fingers when one places ones index fingers comfortably on the indentations on the “F” and “J” keys) are more likely to be pressed as a participant randomly wiggles his or fingers. Further, when a participant randomly presses keys on a keyboard, s/he is more likely to press keys that are close to each other. Therefore, entries that contain letters which are far apart from each other on a keyboard may be more likely to be valid.

I will describe these variables using standard computer science terminology. *QWERTY* is a sort of keyboard layout for Latin script. It refers to the way that letters, other symbols, and functions, are divided amongst the keys on a keyboard. The letters Q, W, E, R, T, and Y are the first 6 letters on the top row from the left. Most keyboards have this layout (Noyes, 1983). A *character* is a printed symbol. Letters, numbers, and punctuations marks are characters, but there are also special characters that one can produce with a QWERTY keyboard such as &, # and %. An *n-gram* is a combination of characters. “A” is a one-gram, and “aj” is a two-gram. An n-gram has a fixed length of n. A *string* is a sequence of characters, regardless of whether the string has an established meaning or not. “Happy” is a string, and so is “nlewr 9h@*Jn l.” A string can

grow longer or shorter as a participant adds or deleted characters from it.

Compositional Variables

The *One-Gram Frequency* variable quantifies how likely it is for any one of the characters that compose a string to come from a valid nonword. Some one-grams are used more frequently in English texts than other one-grams (Norvig, 2009). The most frequent letters are exponentially more frequent than infrequently used letters. This is likely to be true of valid entries as well. However, frequency distributions of letters for the nonwords that participants invent may be different from the frequency distributions of letters in English words, so it is important to compute the frequency with which letters occur using a corpus of invented words. The frequency distribution of letters in invalid entries is likely to be different from this. I hypothesize that letters that occur infrequently in nonwords, but which are close to home position, are more likely to appear in invalid strings.

To calculate the One-Gram variable, I first determined how many times any one letter occurred in the strings produced in the valid condition and divided this count by the total number of letters contained in these strings. Each letter has a frequency score associated with it. I produced another dictionary of frequency scores for entries from the invalid condition. To use these dictionaries to produce the one-gram variable, I first broke up each string of interest into its constituent letters. Each letter got two scores: a score for how frequently it occurred among entries produced in the valid condition, and a score for how frequently it occurred among entries produced in the invalid condition. For each

condition in turn, I took an average. Subtracting the second average from the first average yielded a variable that is positive when the string has one-grams which occur more frequently in the valid condition, and is negative when the string has one-grams which occur more frequently in the invalid condition.

The *Two-Grams* variable quantifies how likely it is for adjacent two-letter combinations to come from a valid nonword. For example, the two-grams which compose “frate” are “fr,” “ra,” “at” and “te.” The One-Grams and the Two-Grams variables are not redundant with each other. Some single letters occur frequently (e.g., “k” and “t”), but rarely occur together (e.g., “kt”).

The Two-Grams variable is computed analogously to the One-Grams variable. There are 676 two-grams in the Latin alphabet (i.e., 26 letters * 26 letters - from “aa” to “zz”). Of 676 possibilities, more than 500 occurred at least once in the invalid condition. However, of the same 676 possibilities, over 250 never occurred once in the valid condition. It is implausible for many possible two grams, such as “kt” to be used in a valid nonword because it is unpronounceable.

The *Vowel and Consonant Profile (VCP)* variable is defined in terms of consonant and vowel sequences within a string. Most strings are composed of substrings which are groupings just of vowels, or just of consonants. For example, “frapjoberyn” has one three-consonant grouping, “ryn,” two two-consonant groupings, “fr” and “pj,” and three single-vowel groupings, “a,” “o” and “e.” The VCP of a string is the number of one, two, three, or four-or-more groupings of vowels and number of one, two, three, or four-or-more groupings of

consonants that compose it. This profile is likely to be different for valid and invalid strings. Invalid strings are likely to have more unusual profiles.

I computed the Vowel and Consonant Profile variable in a similar way to how I compute the One- and Two-Grams variables. That is, I first determined the VCP for each entry in the valid condition. Then I determined the frequency with which any one VCP profile occurred for strings of the same length. I develop the VCP dictionary for strings from the invalid condition in the same way. Entries longer than 9 characters were parsed into shorter strings because there is not sufficient data to meaningfully determine the frequency with which any one VCP profile longer than 10 characters is produced in either of the conditions. Then, after compiling the VCP dictionaries, I computed the VCP for each string of interest. Each string got two new scores: one reflecting the frequency of its VCP among valid strings of its length, and one reflecting the frequency of VCP among invalid strings of its length. If the VCP of a string did not occur in one of the dictionaries, it got a frequency score of 0 for that dictionary. Subtracting the latter score from the former score yielded a variable that has positive values when the strings have a VCP profile that more resembles valid strings, and is negative when it more resembles invalid strings. Strings longer than 10 characters that are broken into substrings still have only one final score which is determined by taking the average of the VCP difference scores of its substrings.

The *Long Consonant Sequences* variable stands for the total number of consonants in sequences of consonants that are longer than four letters. The Long Consonant Sequences variable heavily taxes strings that have

unpronounceable components. This variable is not redundant with VCP. VCP ascribes all sequences of four or more consonants the same value.

The *Invalid Characters* variable is a simple count of characters in a string that do not typically occur in English words. Instructions in all studies clearly indicated that participants are to use only letters (i.e., a-z) to write their invented words. Further, most non-letter symbols on the QWERTY keyboard do not have standard pronunciation when embedded into words. For these reasons, it is likely that most entries that contain non-letter characters are produced by random pressing of characters. Some English words contain hyphens and apostrophes. Some participants might choose to punctuate their entries in a manner that is consistent with English traditions for punctuation. Many English phrases contain space (e.g., “green beans,” “a lot”). For these reasons, the Invalid Characters variable does not count the number of hyphens, apostrophes, periods at the end of entries, or spaces that occur in strings.

Production Variables

The *X-Axis Variability* variable quantifies how spread apart the characters that compose a string are from the left to the right side of the QWERTY keyboard. The spacial distribution of the characters which compose a string is an index of the effort that participant exerted as they produced it. Participants who randomly wiggle their fingers a few times as their hand(s) hover(s) above the keyboard are likely to hit keys that are close to each other in QWERTY space. Participants who first invent a word in their mind and then go to type it select

characters based on their linguistic attributes and not on their spacial distribution, and thus more likely to select characters which are more spread apart.

In order to compute X-Axis Variability, I first measured the distance between keys on several QWERTY keyboards manufactured by different companies. I averaged across these measurements in order to derive an estimate of the distance between keys on the “ideal” keyboard. The tilde key (i.e., “~”) was ascribed a positional value of 0, and each number key to the right of that was ascribed a positional variable relative to that. Our coordinate system integrates the fact that lower rows (i.e., the rows starting with “Q,” “A,” and “Z”) are not directly aligned with the first row. Spaces that are entered as part of strings which have characters in them that are on the right side of the keyboard have a positional coordinate slightly below the “N” key because when people press a space bar they usually do so with their right hand. As an exception, spaces that are part of strings which are written entirely with the left hand are given a positional coordinate a little below the “V” key. After ascribing a positional coordinate to each character, I decomposed each string into its constituent characters and calculated the distance between each character along the x-axis. Finally, I divided this sum by the total number of distances between the characters that make up a string. For example, the string “luzy” has 6 distances: l-u, l-z, l-y, u-z, u-y, and z-y.

The *Y-Axis Variability* variable quantifies how spread apart the characters that compose a string are from the top to the bottom of the QWERTY keyboard. Its rationale is similar to the rationale of the X-Axis Variability variable, and it is

computed in an analogous way. Note, for example, that fingers wiggling in the home position will have very low scores if most of the keys they hit are along the row starting with "A." The X- and Y- Axis variability variables are computed independently of each other, rather than being integrated into a Euclidian distance variable because variability along the x and y axes of the QWERTY keyboard might contribute to differing degrees to the differentiation between valid and invalid entries.

The *Hand-Specific Total Variability* variable stands for the total distance between the characters of a string that are on the left side of the keyboard added to the total distance between the characters of a string that are on the right side of the keyboard. Individuals who produce strings by randomly pressing keys with both hands might have smaller hand-specific variability.

To compute the Hand-Specific Total Variability variable I first decomposed a string into its constituent characters and determined whether each character is on the left or right side of the keyboard. The left side of the keyboard is defined as the keys "5," "T," "G" and "B" and all of the keys to the left of them. All keys that are not left-side keys are right-side keys. Then, I determined the Euclidian distance between each two-key combination for the keys from each side of the keyboard. I divided the sum of all of these distances by the total number of distances for each hand. For example, "luzy" is composed of the left side characters "z" and the right side characters "l," "u" and "y." There are no distances to compute for the left side of the keyboard, and three distances to

compute for the right side of the keyboard. Finally, I add together left hand and right hand specific average distances.

The *One-Handedness* variable specifies whether all of the characters that compose a string come solely from the left side or the right side of the keyboard. Participants who rush through an experiment to finish as quickly as possible might position one hand on the mouse to press the button on the screen which allows them to navigate through the experiment and their other hand in home position quickly pressing keys before moving on. This variable is not redundant with the other procedural variables because it is possible to produce a lot of variation in distance along the axes by using just one hand. The One-Handedness variable is computed straightforwardly by determining whether all of the characters that compose a string are from the left side or from the right side of the keyboard. Entries that are composed entirely of characters from one side of the keyboard receive a 1 for this variable. All other entries receive a 0.

Modified Real Words

Entries that are mere English words or mere modifications of English words should be dropped. For example, although “computery” is an invented word, it straightforwardly calls to mind the word “computer.” If a participant would invent this word it would probably mean “computer-like.” Entries are also invalid if they contain slang words, onomatopoeias (e.g., “boing”) or interjections (e.g., “aha”) because these also have culturally-derived meaning.

In order to compute the Modified Real Word (MRW) variable, I determined the percent of any given entry that is composed of real English strings. An

algorithm written in the programming language Python is used to select the combination of substrings that accounts for the greatest percentage of any given string. For example, the set of substrings (chalk; a; line) accounts for a greater percentage of the entry “chalkaline” than the set of substrings (chalk; line). The algorithm only selects substrings that account for non-overlapping portions of the full string. Thus, the subset (chalk; alkaline) cannot account for the entry “chalkaline” because both “chalk” and “alkaline” require the letters “alk,” and the letters “alk” only occur once in the string.

Entries that can be fully accounted for by English words are not considered to be valid invented words, and are therefore dropped. A more conservative criterion would not be fair to participants because many invented words contain shorter words by pure chance. More than half of entries from the main study have at least one real English word embedded in them. Most of these English words are probably embedded by accident because many three-letter strings are also English words. In fact, even some entries that are composed entirely of English words are probably valid (e.g., “chalkaline”). However, I opted to drop all entries that are perfect matches in order to avoid ad-hoc decisions.

I compared all substrings embedded within an entry with a dictionary that is composed of several sub-dictionaries. The English dictionary included approximately 60,000 frequent English words compiled by Baayen, Piepenbrock and van Rijn (1993). Including only frequent English words helped ensure that I not drop entries that participants invented not knowing that they are, in fact, an obscure component of the English lexicon. The onomatopoeia (“English

onomatopoeias," 2014) and interjection ("English interjections," 2014) lists dictionaries were compiled from Wiktionary, an online open-source dictionary. The slang dictionary was derived from urbandictionary.com, an open-source internet repository of slang terms and definitions. Specifically, I included all of the most popular slang entries for each letter, a-z (e.g., "Most popular words in A," 2014). Open sources for slang terms are likely to be superior to scholarly compilations. Slang quickly changes in time - faster than experts can document. Urbandictionary.com is likely the best repository for the current terminology that participants in our studies, mostly college-aged students, use. In addition, I included a thorough list of contractions in English to handle the fact that many English words contain apostrophes ("List of contractions," 2014).

Some combinations of real English words are held together by short words such as "a." For this reason, our algorithm included a few short single letter words and some of the most common two-letter words. A potential problem with this inclusion is that it could lead to an over inflation of the real word modified variable because many perfectly valid invented words contain these very common single-letter words. To counteract this threat, the algorithm has the additional constraint that single letter words can only count towards the MRW variable if the string also contains three other strings at least three characters long, two other strings at least four characters long, or one other string at least five characters long. This makes it more likely that single character words are intended as stand-alone words. Another source of error for the MRW variable is the fact that English contains many obscure three letter words. In order to ensure

that valid entries are not dropped due to chance matching, the algorithm contains the additional constraint that strings that are three characters or shorter cannot count towards the MRW variable if they are embedded in strings of seven characters or less unless they appear in conjunction with a match that is at least four characters long.

Results of the Validity Study

The analyses were carried out using MPlus, version 7.3 (Muthén & Muthén, 2012). MPlus uses a Bayesian estimator for cross-classified models. Bayesian estimating algorithms do not require that residuals be normally distributed, as traditional estimation algorithms do.

Several of the variables were transformed. Logistic regression, like multiple regression, assumes that predictors are linearly related to the logit of the dependent variable. All the transformations ensured this linearity. Linearity was ascertained for the model that included random slopes for all variables as well as for the final model. Transformations also ensured that all the standard errors of all variables are of the same order of magnitude, roughly between 0.1 and 1.0.

All entries with non-zero scores for the invalid characters variable were produced in the invalid condition. This variable is therefore inappropriate for logistic regression. Variables that perfectly predict group membership are known as complete separators. Complete separators are anomalous in that their fixed effects and standard errors are often immense. It is therefore standard practice to exclude such variables from logistic regression. Although information regarding the invalid characters variable will not be formally modeled using the cross-

classified analysis below, I will make use of knowledge garnered by it by dropping all future entries that contain any invalid characters.

In the first series of models, the relationship between most predictors and the dependent variables was modeled as fixed. Each of the models in the first series had a different slope modeled as random. For all models, random effects were significant at both upper levels at the .05 level. The second model included random slopes for all predictor variables. All random slopes, when tested one at a time, were significant at both upper levels.

After determining that the relationship between each predictor and the dependent variable is dependent upon both picture condition and participant, I created a model that allowed all variables to have random slopes across both upper cross-levels. Then, I trimmed this model down in two phases. In phase one I dropped variables that had fixed effects that were not theoretically sensible. Fixed effects which are in the wrong direction are not to be trusted. It is also relevant to note that none of these unpredicted fixed effects were statistically significant. After dropping each variable I computed each model again before again dropping variables with theoretically implausible signs. The variable with the largest fixed effects that was not theoretically sensible was Hand-Specific Total Variability. After dropping it, one more variable had theoretically senseless fixed effects: x-Axis Variability. In the second phase of the trimming process, I dropped variables with fixed effects that were not significantly different from 0. Again, I recomputed models after dropping each variable.

The final model is composed of variables that have significant main effects, significant fixed effects, and theoretically meaningful fixed effects. The model is $\text{logit_final} = \log(\pi/(1-\pi)) = -0.198 + (-1.374 * \text{Long Consonant Sequences}) + (\text{Two-Grams} * 6.552) + (\text{Vowel and Consonant Profile} * 2.172)$. The significance of each of these fixed effects is less than 0.000. Pi, or the probability that a given case comes from the invalid condition is $1 / (1 + \exp(-\text{logit_final}))$. This model correctly determines the condition in which strings were produced with 79.4% accuracy.

Results of the Main Study

After eliminating gibberish entries, I built a structural equation model to assess whether the emotional properties of pictures predict the emotional properties of invented words. A structural equation model has two components; a measurement model and a path analysis. A measurement model is used to examine whether the variables of interest are measured correctly. In the current study, it is used to assess whether nonword coders were able to reliably assess nonword properties. Determining whether coder ratings cohere in a meaningful way is necessary in order to establish that nonword ratings meaningfully reveal nonword valence and nonword arousal. Coherence is assessed by examining the fit of models that are composed of both latent and observed variables. Latent variables are constructed by examining patterns of covariance among observed variables. Observed variables are direct measurements. Latent variables are the abstract concepts one attempts to indirectly index through these direct measurements. In the current study, individual coder ratings of valence and

arousal are observed variables. Valence and arousal connotation of invented words are modeled as latent variables. One of the benefits of modeling the emotional connotations of invented words as latent variables, rather than merely averaging across coder ratings, is that one can now partition true measurement (signal) from method error (noise).

Measurement models can be constructed through either confirmatory or exploratory factor analysis (CFA and EFA, respectively). CFA is used when one has a theoretically-derived hypothesis of what the ideal measurement model is. EFA is used when one does not have a clear hypothesis and needs an atheoretical method to derive an acceptable measurement model. Most of the methods used to produce estimates for CFAs and EFAs allow one to formally assess how well these models fit the data. This is done through fit indexes. A thorough description of how one can use these fit indexes can be found in Bollen (1989). I therefore use CFA techniques to examine whether valence and arousal latent variables best account for coders' ratings. I use EFA to examine problematic features of the models that I did not expect.

One constructs a measurement model before constructing a full structural equation model. A structural equation model is an examination of how it is that latent variables relate to each other. It is necessary to construct a measurement model before doing this because one needs to ascertain that concepts are measured correctly before determining how it is that these concepts correlate. The path analysis for this study needs to be a cross-classified structural equation model. Therefore, the measurement model also needs to be cross-classified.

Unfortunately, this is not possible. Cross-classified analysis in MPlus uses a Bayesian estimator. This sort of estimation does not typically produce fit indexes that maximum likelihood procedures produce. Even though it is not possible to directly assess the cross-classified model fit, it would be informative to assess model fit ignoring nesting and then for a multilevel model that has only one second cross-level.

In order to construct these measurement models, I first parceled coder responses. Parcels are a type of observed variable used to measure an underlying latent variable. They are constructed by averaging several responses together. Parcels are different from latent variables. One creates parcels by averaging the scores of several coders. Each coder's score receives equal weight. Latent variables, on the other hand, represent theoretically meaningful shared variance among multiple measurements. Even though a latent variable is technically a linear composite of its various observed variables, each observed variable received a different empirically-determined weight. The ratings among the observed variables of a latent variable can be allowed to correlate. While the latent variable represents theoretically meaningful shared variance, correlated error terms represent theoretically meaningless methodological noise. Coders were grouped in the same way for the valence and arousal parcels, so in the models below I allow the error variance of each parcel created for the valence latent variable to correlate with the analogous parcel used to create the arousal latent variable.

There are several different ways to construct parcels (Little, Cunningham, Shahar, & Widaman, 2002). In the current context, there is no intrinsic reason to group any two coder's responses together. Therefore, I grouped coders into parcels through a random process using SPSS's random number generators. These parcels were, in turn, used as observed variables to model nonword valence and nonword arousal as latent variables. It is ideal to model latent variables with three observed variables. Two of the parcels therefore contained four ratings, and the third parcel contained five ratings. A CFA that did not account for nesting revealed that this model had less-than-ideal fit (RMSEA = 0.128, CFI = 0.909, TLI = 0.829, Chi-Square = 442.081 with 8 degrees of freedom). A series of exploratory factor analyses were carried out in order to identify the problem. I examined the fit indexes for the two-factor solution for a model with all coders, and then a model with dropping out each coder at a time and keeping all the other coders in. After excluding the coder who accounted for the greatest decrement in the EFA fit indexes, the four parcel CFA (now with four coders per each parcel) had acceptable fit (RMSEA = 0.093, CFI = 0.952, TLI = 0.910, Chi-Square = 239.422 with 8 degrees of freedom). One can therefore think of this coder as an outlier. His or her responses were different enough from the average response that these should be eliminated from further analyses so not to risk introducing unnecessary error in the measurement model.

After dropping data from this coder I next tested a multilevel CFA with observations nested within pictures. A multilevel CFA is used to determine whether the latent variable that is being measured should be measured in the

same way for different levels. Latent variables were modeled at both levels 1 and 2 with the constraints that loadings are identical at both levels and that factors correlated to the same degree at both levels. This model had an even better fit than the model which did not take nesting into consideration (RMSEA = .069, CFI = 0.944, TLI = 0.920). However, this model suffered from an estimation problem in that one of its error variances was slightly negative. Constraining this error variance to be zero did not worsen model fit. The cross-classified measurement model is therefore probably also valid.

The next step after constructing a measurement model is to examine how it is that variables of interest correlate. In the current situation I am interested in how it is that picture emotional features correlate with nonword emotional features. To assess this, picture valence and picture arousal were entered as between-picture variables because different pictures have different emotional properties. Both nonword valence and arousal were regressed on both picture valence and arousal. These four correlations are entered in the same model at the same time.

Picture valence predicted nonword valence ($r = .082$, 95% CI is 0.056 to 0.110), but not nonword arousal ($r = -.022$, 95% CI is -0.054 to 0.007); picture arousal predicted nonword arousal ($r = .056$, 95% CI is 0.032 to 0.082) and also nonword valence ($r = -0.020$, 95% CI is -0.040 to -0.001). This model is depicted in Figure 1. It is interesting to note that although this last effect is statistically significant, picture valence predicted nonword valence more than it did nonword arousal, and picture arousal predicted nonword arousal more than it did nonword

valence. This double dissociation is to be expected in light of the core theory of affect, which holds that valence and arousal are orthogonal.

A Confound Analysis

It is possible that the only reason invented words mediated affect is that these invented words resemble English words that have a clear emotional meaning. Entries might be composed of real English words. It is also possible that an entry as a whole resembles an English word. These are two separate levels of analysis and each requires a different approach.

In order to account for the possibility that invented words connote affective information due only to embedded emotional English words, I first examine all of the English words that are embedded within any given string. This method is similar to the method that was used to determine whether any given string is composed entirely of English words. The difference is that whereas the previous approach involved determining the combination of English words that best accounted for the greatest percentage of the entry, this approach considers all substrings of a given string which are English, regardless of whether these substrings account for the same letter of the string or not. The only constraint is that embedded substrings are longer than two characters. For example, this approach will consider all of the substrings of the string “coolanda,” including “coo,” “and,” “cool” and “land.” This approach is ideal because it does not require any assumptions about what it is that participants had in mind when producing an entry, and it does not make assumptions regarding which features the entry coders might notice.

The second approach to account for the alternative explanation that invented words convey emotional information due only to their resemblance to English words that have clear emotional denotation involves determining the real English words which are most similar to invented words. This approach is inspired by the work of Yarkoni, Balota and Yap (2008) who found that a word's 20 nearest Levenshtein distance neighbors influence how it is that word is perceived. The Levenshtein distance between two strings of letters is the number of changes that need to be made for one string of letters to become totally like another string of letters. The three types of changes that are needed are adding letters, dropping letters, or swapping letters. For example, the string "radzer" is a Levenshtein distance of 2 away from the string "radar" (radzer -> (drop the z) rader - > (swap the e with a) radar). The package "vwr" (Keuleers, 2013) was written for the statistical program R, version 3.1.0 (R Development Core Team, 2013) to compute the Levenshtein distance between any two strings. vwr requires a list of sources (in the current case, a list of invented words), and a list of targets (in the current case, a list of English words). I used vwr to compute the 30 closest English emotional words to any given invited words. Levenshtein distance is computed in counts of natural numbers, so a word can be at a Levenshtein distance of 1, or 2, or 3, etc. I took all words that are at the smallest Levenshtein distance from each invented word. For "radzer," these are the words badger, ladder, madder, radar, rapier, rather, razor, reader and rider (I did not find English words with at a Levenshtein distance of 1 from "radzer"). If the very

few instances that an invented word had more than 30 neighbors at its nearest distance, I took the first 30 in alphabetical order.

A dictionary of English words with emotional meanings was compiled by combining four separate dictionaries. The Affective Words for English Words (ANEW; Bradley and Lang, 1999) contains words that are continuously rated for valence and arousal. The Dictionary of Affective in Language (DAL; Whissell, Fournier, Pelland, Weir and Makarec, K, 1986) has continuous ratings for pleasantness and activation which are akin to valence and arousal. The Linguistic Inquiry and Word Count (LIWC; Pennebaker, Francis and Booth, 2001) and the General Inquirer (GI; Stone, Dunphy, Smith & Ogilvie, 1968) contain lists of English words that are categorized discretely as related to positive emotions, negative emotions, or not related to emotions. The LIWC has two sorts of entries: words and stems. Stems are letters that have multiple possible endings that are emotionally similar (for example, “admir-” can be expanded to “admiration” “admirer,” etc). I expanded LIWC stems into all entry matches of the same valence from the other dictionaries. When a stem did not have a match in the other dictionaries, the stem was either a stand-alone emotion word and thus did not need further expansion, was expanded into all possible extensions contained in the 60,000 word version of the CELEX lexical database (Baayen, Piepenbrock & van Rijn, 1993), or was expanded with the help of Google. The GI contains multiple definitions for many of its words. In the case that the multiple definitions had different emotional connotations, I included the definition that is most frequently used in English. Only 18 words from the GI both a positive and a

negative meaning. In all of these cases, one of the emotional meanings occurs less than 10% of the time.

The ANEW has 1034 emotion entries, the DAL has 8742, the LIWC 905, and the GI 3452. The combined dictionary has 11495 entries. Any two dictionaries share enough terms for correlations to be computed among the dictionaries and for a CFA to be constructed. Valence and arousal entries from the ANEW and the DAL were entered as continuous observed variables. The error variance of the valence and arousal variables from each of these dictionaries were allowed to correlate to allow for shared method variance. The LIWC and GI were combined into a single dichotomous observed valence variable. The LIWC and the GI had the same valence score for all but around 50 entries. These 50 or so entries had ambivalent meanings and were thus excluded.

I hypothesized that the correlations among these observed variables are due to two underlying latent variables, a valence and an arousal latent variable. In order to identify this model, the variance of these latent variables is constrained to 1. Unsurprisingly, this model has excellent fit (RMSEA = 0.023, CFI = 0.996 and TLI = 0.982). This model is depicted in Figure 2. Although a lot of data is missing (most entries did not have scores for all three valence variables) it is likely that data is missing at random. Therefore the WLSMV estimator used by MPlus to estimate this model is appropriate. The CFA can be used to produce valence and arousal factor scores for all of the entries in the final dictionary. Factor scores are composed by considering both the scores that

each entry has for the various dictionaries that it happens to be part of as well as the loading that observed variables have onto the latent variables. Factor scores are theoretically grounded, are more reliable than mere averages would have been, and are robust to the fact that most words do not appear in all of the dictionaries that we used.

The final dictionary can be used to compute two of the possible confounding variables. First, after decomposing each entry into all possible substrings that are emotion words, I can take the average valence and average arousal factor scores of these substrings. Second, I can take the average of all closest Levenshtein neighbors to each entry. These two variables are different enough not to be collapsed, but similar enough to warrant inclusion in the same statistical model. This statistical model is very similar to the cross-classified measurement model/path analysis constructed for the main study. The only difference is that emotion word variables are now added as covariates. Specifically, the latent variable for nonword valence is now regressed onto the average valence of its 30 closest Levenshtein neighbors, the average valence of all embedded real words (both of which are modeled as observed variables) as well as onto picture valence and picture arousal. Further, the latent variable for nonword arousal is now regressed onto the average arousal of its 30 closest Levenshtein neighbors, the average arousal of all embedded real words (both of which are modeled as observed variables) as well as onto picture valence and picture arousal. If the direct effects between pictures and nonword emotional features subsist after entering the covariates that are relevant for each latent

variable, then these covariates are not actually confounds. The results are that picture valence still predicts nonword valence and picture arousal still predicts nonword arousal after considering the potential confounds. For valence, the correlation is slightly stronger ($r = 0.110$, $p = 0.000$), and for arousal the correlation is slightly weaker ($r = 0.044$, $p = 0.050$). This model is depicted in Figure 3. Invented words carry affective information in spite of, not because of, the emotional connotations of the English words they resemble.

Discussion

Summary

The nonwords that people invented had emotional undertones which reflected the emotional undertones of the pictures that the invented words referred to. This finding demonstrates both that English speakers can encode the meaning of their emotional experiences into invented words, and that other English speakers can decode the meaning of these invented words. While it has already been shown that people can decode the meaning of words that they have never been exposed to before (from both their own language and languages that they have never heard), this is the first time that it has been demonstrated that people can meaningfully invent words to stand for their emotional experiences.

These effects are not due to the presence of two kinds of invalid data: entries produced by randomly pressing keys on a keyboard, and entries that resemble English words too much. Further, these effects subsist even after accounting for the possibility that invented words transmit affective information

due solely to their resemblance to English words that have emotional meanings.

Conclusions

These findings have implications for both dimensional views of emotion and for theories of emotional linguistic relativism. First, they make it plausible that the same important dimensions which define the abstract semantic space within which emotion words are organized also define the abstract semantic space within which invented emotion words are organized. This is a strong argument in the favor of the view that the same basic dimensions underlie all emotional experience. Future studies should assess the regions of this abstract space that invented words populate, and test whether they tend to fall in regions within which human emotional experiences tend to lie.

Although the current findings support dimensional theories of emotions, they do not as clearly support linguistic relativism. That English speakers can decode the meaning of invented words without knowing anything about the context within which words were invented complicates the statement that “the meaning of invented words is culturally determined” in several ways. First, we introduce, and empirically examine, a new type of utterance that proponents of linguistic relativism have not debated as comprehensibly as they have debated natural language (i.e., words used in natural language). It is possible that the meaning of invented emotion words is not culturally determined, or at least not all the way. This is demonstrated by the fact that English speakers can reliably decode the valence and arousal connotations of invented words without ever having been exposed to these invented words before and without knowing

anything about the context within which they were produced.

Second, these findings should shift the debate of linguistic relativism from the word to subword sounds. It is likely that sound symbols which occur in invented words account for the transmission of meaning. How it is that these more atomic units of meaning account for the transmission of affect is left for future studies to assess. If the “soft” interpretation of the sound-symbolic theory is true, this would be consistent with a liberal interpretation of linguistic relativism. A linguistic relativist can accommodate our findings by claiming that the meaning of subword units is culturally derived. If the “hard” interpretation of the sound-symbolic theory is true, this would be highly problematic for linguistic relativism. It would offer a mechanism to explain how it is that the meaning of some words is intrinsically linked to the sound of the word.

Future Direction: Why Study Invented Language?

It is important to study how people invent words because this offers important insight into the basic dimensions which underlie meaning production, and because it produces important information worth considering while adjudicating between theories concerning the relationship between language and meaning. In addition to these theoretical benefits, studying how people invent words may confer a number of methodological and practical benefits.

Value to Researchers: Measuring Emotions

Attitude (McConnell & Leibold, 2001), motive (Pang, 2010), and to some degree emotion (Barrett, Niedenthal & Winkielman, 2005) researchers posit that there are two types of systems: explicit and implicit. Explicit attitudes, motives,

and emotions are assessed by directly asking a person what they think, want, or feel. These are mental states that a person can verbally report. Implicit attitudes, motives, and emotions are difficult to verbally report. For this reason researchers rely on indirect methods to measure them. Examples from the three domains are the implicit attitudes test (for attitudes), the picture story exercise (for motives) and the implicit positive affect negative affect test (for emotion; IPANAT; Quirin, Kazén & Kuhl, 2009). The IPANAT is indirect in that instead of prompting participants to self-report on their emotional state, it prompts participants to rate a series of invented words for the degree to which they “look like” they connote any one of several emotions. The IPANAT developers theorize that, unbeknownst to participants, the nonwords in themselves have no intrinsic meaning. Thus, the ratings that participants produce reflect the participants’ own implicit emotional states.

Asking participants to invent words might be another good way to measure their implicit emotions. To assess a participant’s implicit emotional state one would need to ask a participant to invent an emotional word to assess a nonemotional stimulus. As the emotional connotations of the invented word do not describe the stimulus, they may be a good indicator of the participant’s implicit emotional state. As the current study only assessed correlations between the emotional connotations of invented words and the emotional ratings of stimuli that were consciously derived, future studies are needed to assess this possibility.

Several features of the word invention task make it a promising

measurement for emotion. First, our measurement is useful for studies in which a large amount of implicit measurements for emotion are needed. In these cases, participants may habituate to traditional measurements for implicit affect faster than they would to the current measurement for affect. Second, our task is highly malleable. It is very easy to come up with cover stories for it. This makes it easy to embed the nonword generation task into a variety of experimental paradigms into which it would have been awkward to embed other implicit measurements of affect. Third, a single nonword can be assessed along at least two dimensions (valence and arousal). It may be possible to pick up on other types of emotional information from invented words. Future studies need to assess this. If this would hold, then the nonword generation task could be a good way to derive many measurements from participants with a single question.

Value for Participants: Explicating Hard-to-Verbalize States

Researchers might benefit from a new way to assess how participants feel. Participants might also benefit from a new way of expressing their emotions. First, inventing words will be useful when a person is in an emotional state that the linguistic community that s/he is part of does not have a word for. Second, inventing words will be useful when a person attempts to explicate complicated states that involve an unusual combination of lower order emotional states. People who have a hard time conceptualizing ambivalent states might only articulate one of their reactions. Similarly, people who abide in societies that enforce consistency norms will feel a pressure to not express contradictory feelings. If a person has the opportunity to invent emotion words he or may

discover a way to speak in terms of complicated emotional blends which his or her society deems uncommon.

There are many situations in which a person would not want to invent emotion words. First, if a person already has an emotion word to adequately signify the emotional state that s/he is in, then inventing a new emotion word to stand for that state is at best redundant. At worse, it will deplete the person of cognitive resources and lead to unnecessary confusion. Second, a person would also not want to invent an emotion word if this will lead to interpersonal confusion. In an environment where the stakes of understanding what another person is trying to say are high, it might be worth sacrificing some validity for the high precision that natural language offers. It is important to keep in mind that several methodological strengths of this study are not available in real-time interactions. First, I used many raters to assess the validity of any one nonword. Second, I considered average correlations between over three thousand nonwords and pictures. What is strength in an empirical study is a weakness in real life. These methods allowed us to gloss over the noisiness embedded in nonword production to catch a better glimpse of the constructs that are theoretically interesting. Future studies will need to directly assess whether any one nonword can lead to effective communication between two human beings.

Coda

It is virtually unheard of in the academe for scientists to study how people invent new words. The closest approximation is linguists who study how people combine old word elements to create words that stand for relatively new

concepts that are very similar to old concepts (Štekauer, 2005). The absence of research isn't surprising. Our society's pressure to conform renders the study of word invention in the best case irrelevant, in the worst case problematic. We need to show its benefits, delineate boundary conditions, teach people how to invent language, and promote a culture that is open to linguistic creativity.

Structural Equation Model for Main Study

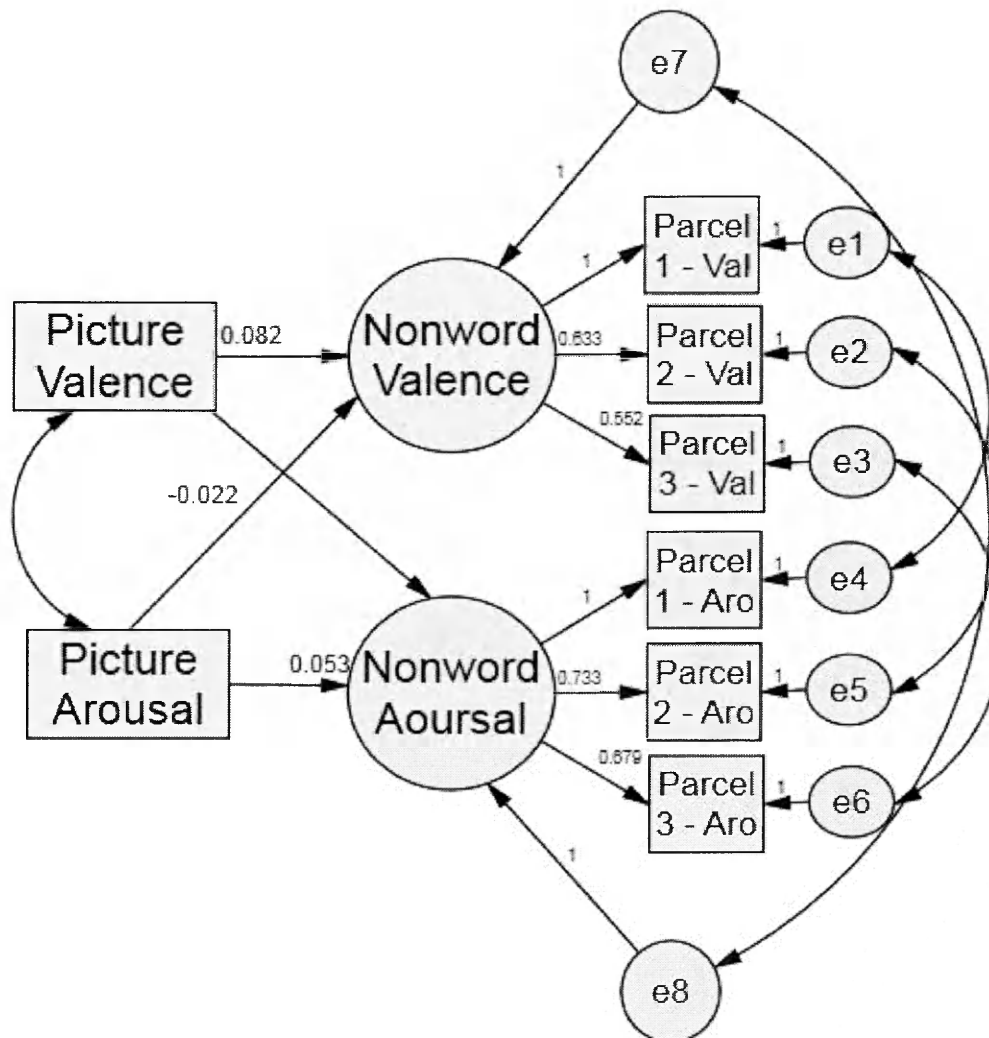


Figure 1. This is a cross-classified SEM. Picture valence and arousal are modeled as level 2 variables (between pictures); all other variables are allowed to have variance at all levels. For this reason, I only depict the level 2 – between pictures SEM.

Nonword valence and arousal are modeled as latent variables reflected by three parcels of coder ratings. Analogous parcels for valence and arousal contain the same coder ratings, and are thus allowed to correlate with each other. Nonword valence and arousal are regressed onto picture valence and arousal. All of the parameters that are included are significant at the .05 level

Measurement Model to Construct the Dictionary Used in the Confound Analysis

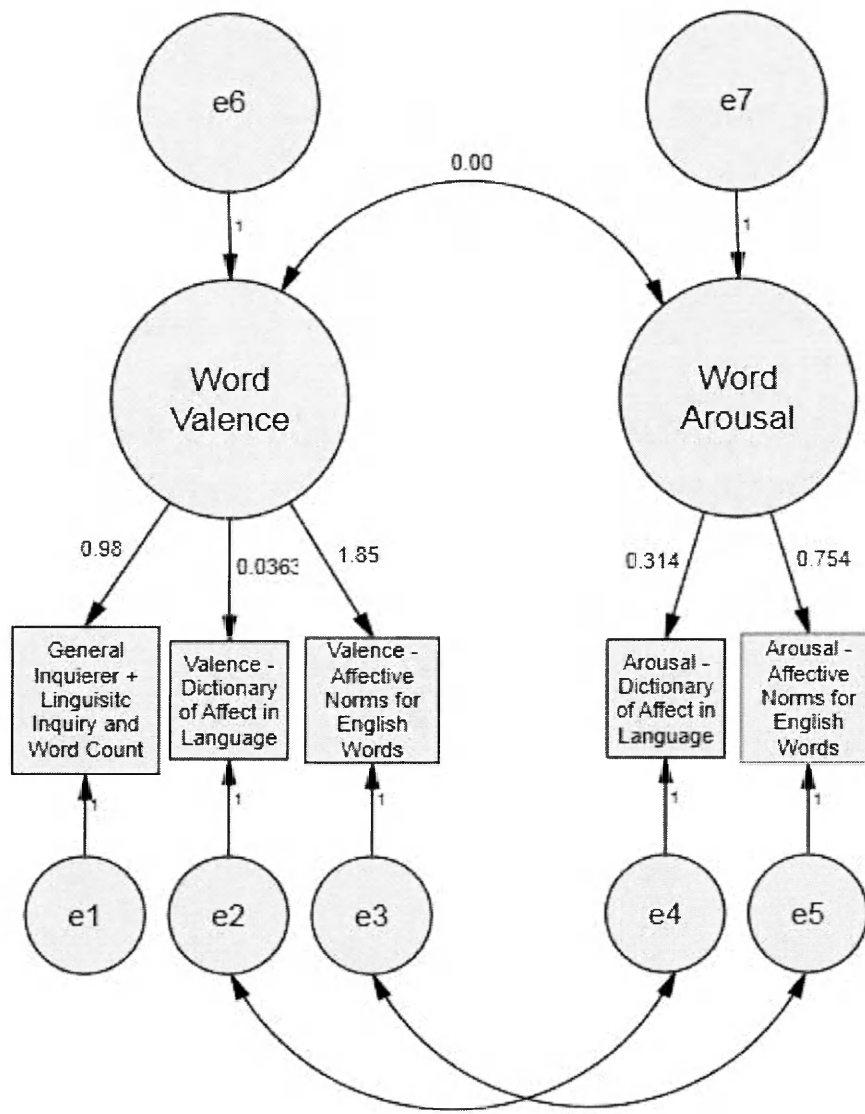


Figure 2. This is a confirmatory factor analysis used to derive valence and arousal factor scores for all of the entries in the final dictionary. The two Dictionary for Affect in Language and the two Affective Norms for English Words observed variables are continuous, while the variable composed by combining the General Inquirer with the Linguistic Inquiry and Word Count is dichotomous.

Structural Equation Model for the Confound Analysis

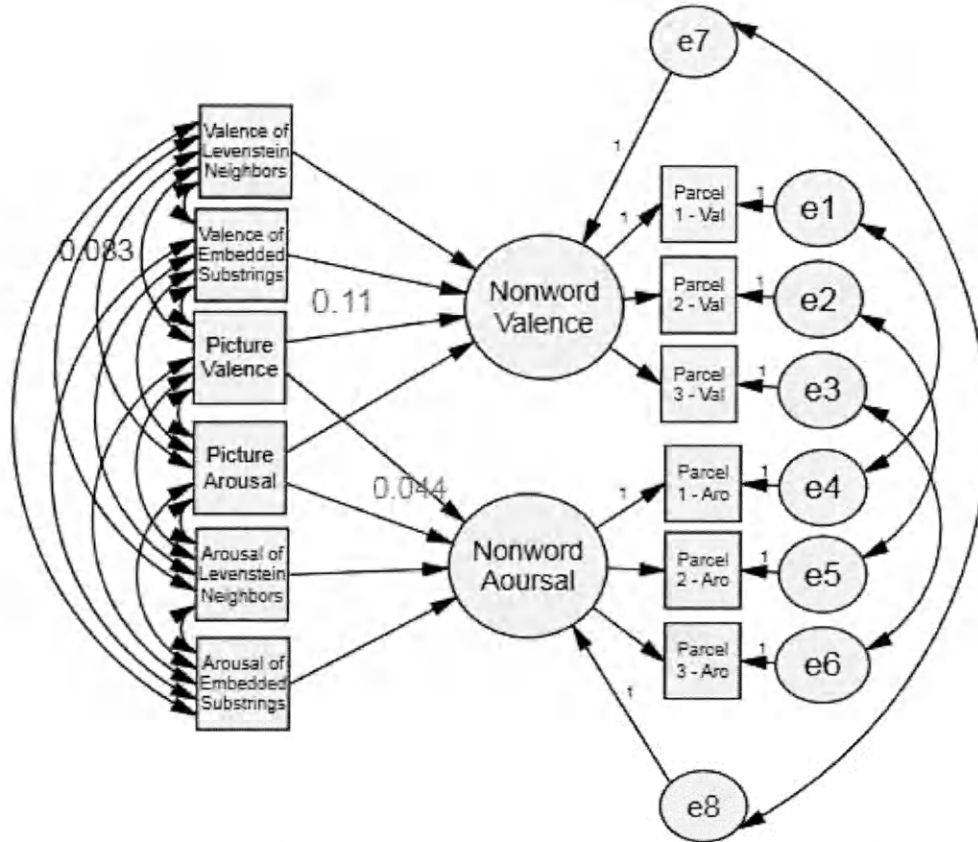


Figure 3. This model is identical to the model in Figure 1, with the exception that additional covariates are regressed onto the Nonword Valence and Nonword Arousal latent variables. These additional covariates account for two different ways in which nonwords resemble English words with known emotional connotations. They are only allowed to have variance at the between-picture level. All of the included parameters are significant at the .05 level.

References

- Barrett, L. F., Lindquist, K. A., & Gendron, M. (2007). Language as context for the perception of emotion. *Trends in Cognitive Sciences, 11*, 327-332.
- Barrett, L. F., Niedenthal, P. M., & Winkielman, P. (Eds.). (2005). *Emotion and Consciousness*. New York, NY: Guilford Press.
- Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). The Celex lexical database (CD-ROM). University of Pennsylvania, Philadelphia: Linguistic Data Consortium.
- Bollen, K. A. (1998). *Structural Equation Models*. New York: Wiley.
- Bradley, M. M., Codispoti, M., Cuthbert, B. N., & Lang, P. J. (2001). Emotion and motivation I: Defensive and appetitive reactions in picture processing. *Emotion, 1*, 276.
- Bradley, M. M., & Lang, P. J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings (pp. 1-45). Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.
- Ekman, P. (1992). Are there basic emotions? *Psychological Review, 99*, 550-553.
- English interjections. (n.d.) In Wiktionary. Retrieved January 20, 2014 from http://en.wiktionary.org/wiki/Category:English_interjections.
- English onomatopoeias. (n.d.) In Wiktionary. Retrieved January 20, 2014 from http://en.wiktionary.org/wiki/Category:English_onomatopoeias.
- Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a

- pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, 103, 54.
- Hackett, C. F. (1960). The origin of speech. *Scientific American*, 203, 88-96.
- Keuleers, E. (2013). vwr: Useful functions for visual word recognition research. R package version 0.3.0. <http://CRAN.R-project.org/package=vwr>
- Koriat, A. (1975). Phonetic symbolism and feeling of knowing. *Memory & Cognition*, 3, 545-548.
- Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1999). International affective picture system (IAPS): Instruction manual and affective ratings. The center for research in psychophysiology, University of Florida.
- List of contractions. (n.d.). In Wikia Retrieved Jun 7, 2014 from http://grammar.wikia.com/wiki/List_of_contractions.
- Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling*, 9, 151-173.
- McConnell, A. R., & Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology*, 37, 435-442.
- Most popular words in A. (n.d.) In UrbanDictionary. Retrieved June 7, 2014 from <http://www.urbandictionary.com/popular.php?character=A>.
- Muthén, L. K., & Muthén, B. O. (1998-2012). Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén.
- Norvig, P. (2009). Natural Language Corpus Data. In Segaran, T., &

- Hammerbacher, J. (Eds). *Beautiful Data: The Stories Behind Elegant Data Solutions* (219-242). Sebastopol, CA: O'Reilly Media, Inc.
- Myers-Schulz, B., Pujara, M., Wolf, R. C., & Koenigs, M. (2013). Inherent emotional quality of human speech sounds. *Cognition & Emotion*, *27*, 1105-1113.
- Noyes, J. (1983). The QWERTY keyboard: A review. *International Journal of Man-Machine Studies*, *18*, 265-281.
- Nuckolls, J. B. (1999). The case for sound symbolism. *Annual Review of Anthropology*, *28*, 225-252.
- Pang, J. S. (2010). Content coding methods in implicit motive assessment: Standards of measurement and best practices for the Picture Story Exercise. In O.C. Schultheiss & J.C. Brunstein (Eds.), *Implicit Motives*. New York, NY: Oxford University Press.
- Parault, S. J., & Schwanenflugel, P. J. (2006). Sound-symbolism: A piece in the puzzle of word learning. *Journal of Psycholinguistic Research*, *35*, 329-351.
- Pennebaker JW, Francis ME, Booth RJ. 2001. *Linguistic Inquiry and Word Count (LIWC): LIWC 2001*. Mahwah, NJ: Erlbaum.
- Quirin, M., Kazén, M., & Kuhl, J. (2009). When nonsense sounds happy or helpless: The Implicit Positive and Negative Affect Test (IPANAT). *Journal of Personality and Social Psychology*, *97*, 500.
- R Development Core Team. (2013). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical

- Computing. Retrieved from <http://www.r-project.org/>
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. (Vol. 1). Sage.
- Russell, J. A., Lewicka, M., & Niit, T. (1989). A cross-cultural study of a circumplex model of affect. *Journal of Personality and Social Psychology*, *57*, 848.
- Stone, P., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1968). The general inquirer: A computer approach to content analysis. *Journal of Regional Science*, *8*, 113-116.
- Weiss, J. H. (1964). Phonetic symbolism reexamined. *Psychological Bulletin*, *61*, 454.
- Whissell, C., Fournier, M., Pelland, R., Weir, D., & Makarec, K. (1986). A dictionary of affect in language: IV. Reliability, validity, and applications. *Perceptual and Motor Skills*, *62*, 875-888.
- Wierzbicka, A. (2009). Language and metalanguage: Key issues in emotion research. *Emotion Review*, *1*, 3-14.
- Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*, 971-979.
- Yik, M., Russell, J. A., & Steiger, J. H. (2011). A 12-point circumplex structure of core affect. *Emotion*, *11*, 705.
- Štekauer, P. (2005). Onomasiological approach to word-formation. In Štekauer, P & Lieber, R (Eds), *Handbook of word-formation* (pp. 207-232),

Netherlands: Springer.