# Evaluation of Linguistic Labels Used in Applications

**Hanan Hibshi** **Travis Breaux**

March 2016
CMU-ISR-16-104

Institute for Software Research
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Institute for Software Research, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA

**Abstract**

Linguistic labels such as high, medium, and low are commonly used in different applications. Researchers in psychometrics argue that before adding new labels to applications, the labels must be empirically evaluated. In this paper, we explain the process of selecting labels for a security assessment application. We also show how we evaluate the labels empirically using a sample population from Amazon Mechanical Turk users.

This research is supported by grants from the National Security Agency and King Abdul-Aziz University.

**Keywords:** Security, requirements, linguistic labels, constructs, evaluation

# 1 Introduction and Motivation

In our efforts to build an application for security experts that helps with security assessments, we found it necessary that our application communicates security levels on some labeled scale. There has been a number of efforts to represent security on a scale, for example, the National Institute of Standards and Technology (NIST) special publication 800-30 recommends three levels to represent security risk: low, medium and high. In this work, we describe how we created new labels to describe security on a scale of adequacy. It is important to realize here that we are creating a new scale to measure our construct of security adequacy, because there are no existing, empirically valid scales to measure this construct. Psychometric researchers describe this type of scale as an ad hoc scale due to the lack of valid or reliable scales [4]. Creating ad hoc scales requires evaluation to examine the reliability of the scales rather than relying on the face validity, alone [4]. On the contrast of construct validity, a face validity is subjective: if a test or measure "looks like" it going to measure what it is supposed measure, then it has face validity [7].

# 2 Research Approach

In this section we describe the details of our research methodology.

## 2.1 Selecting the Initial Label Set

The application that we are building is a security assessment application that relies on linguistic labels that will be modeled as fuzzy sets. The linguistic labels should describe a security assessment metric to users of the system. The choice of such labels is context dependent by application, and relies on the background knowledge supported by empirical evaluation using experiments [5, 6]. We conducted a focus group of five researchers in our lab to discuss the initial set of labels that we are considering for the security assessment application. We used the context of a concrete scenario to be able assess labels that mostly came from prior research in fuzzy logic [5, 6]. An example of some of the labels used in prior research include: low, medium, high, and moderate. The focus group discussion concluded that these labels are not very well suited to describe security, for example: how can we define low security vs. high security?

Experts analyze threats in a security scenario as they are concerned with risks involved with the threat. Security requirements are intended to mitigate the threats and decrease the risk. This understanding of security requirements enables us to describe requirements in terms of "adequacy" of a security requirement to mitigate a threat. For example, a certain requirement can be inadequate, adequate, or excessive.

We choose the three labels: inadequate, adequate, and excessive. Below, we will describe our approach of evaluating the labels experimentally, and the English language proficiency test needed for our experiment results to be reliable.

## 2.2 Experimental Evaluation of the Label Set

As explained above, creating new scales to measure a construct requires proper evaluation that does not rely on face validity alone. We are evaluating three anchor points: inadequate, adequate and excessive. We looked into a standard English dictionary for synonyms of the words: inadequate, adequate and excessive and we created a 17-word data set: the original labels and the additional synonyms. Since these labels will be ranked by participant to describe adequacy, we replace adequate with "average" for the purpose of creating a middle anchor point where we can observe whether the two extremes : inadequate, excessive are actually placed further from average towards on the opposite ends of the ranking scale.

Since humans' perception of adequacy can vary by scenario and context, we chose to create four scenarios where we vary scenarios by adequacy skewness bias. By adequacy skewness bias we mean the bias introduced in a scenario that might cause some bias in adequacy ratings. For example, if we ask participants to rate adequacy while thinking of the size of meal portion at a restaurant, their ratings might be skewed towards inadequacy. In fact, research has shown that it is hard to introduce smaller portion sizes at restaurants to promote health, because of consumer resistance to the smaller portions [1]. Another example is thes U.S. citizens' concerns over their privacy against government surveillance [3]. Below, we list the different adequacy scenarios we showed to consumers:

- **Length of a time waiting for a bus** (skewed towards excessive)
  *Imagine you are waiting for a bus and you want to choose a word to describe the length of time that you have been waiting. Rank the following words from least to greatest as you might use them to describe the wait time*

- **Distance of a walk to a parking spot** (skewed towards excessive)
  *Imagine you are walking from your parked car to a building. You want to choose a word to describe the distance that you have to walk from your parking spot. Rank the following words from least to greatest as you might use them to describe the distance of the walk*

- **Size of meal portions at a restaurant** (skewed towards inadequate)
  *Imagine you sitting down at a restaurant to eat dinner and the waiter brings your food to the table. You want to choose a word to describe the size of the meal portion that you have received. Rank the following words from least to greatest as you might use them to describe the meal portion*

- **Amount of privacy protecting from government surveillance** (skewed towards inadequate)
  *Consider the degree of government surveillance in the United States. You want to choose a word to describe the amount of privacy you have from the government and technology companies. Rank the following words from least to greatest as you might use them to describe the amount of privacy you have*

We recruited participants from Amazon Mechanical Turk (Mturk). Participants were screened for English proficiency (explained in detail later in this section) before proceeding to the word

ranking task. Then, we provided text and video instructions with an example of how to complete the task to reduce ambiguity and learning fatigue/bias. We presented each participant with the four scenarios. The order of the scenarios was randomized and so is the order of the words. Based on pilot results, we estimated the completion time to be 15-25 minutes, so we compensated participants with $3. Throughput the paper, we will refer to this study as the word-ranking study. Figure 1shows an example screen-shot of what was shown to participants. The collected data were analyzed by calculating means of rankings provided for each label. Our goal was to ensure that the labels: inadequate and excessive are listed on the far ends away from the average point.

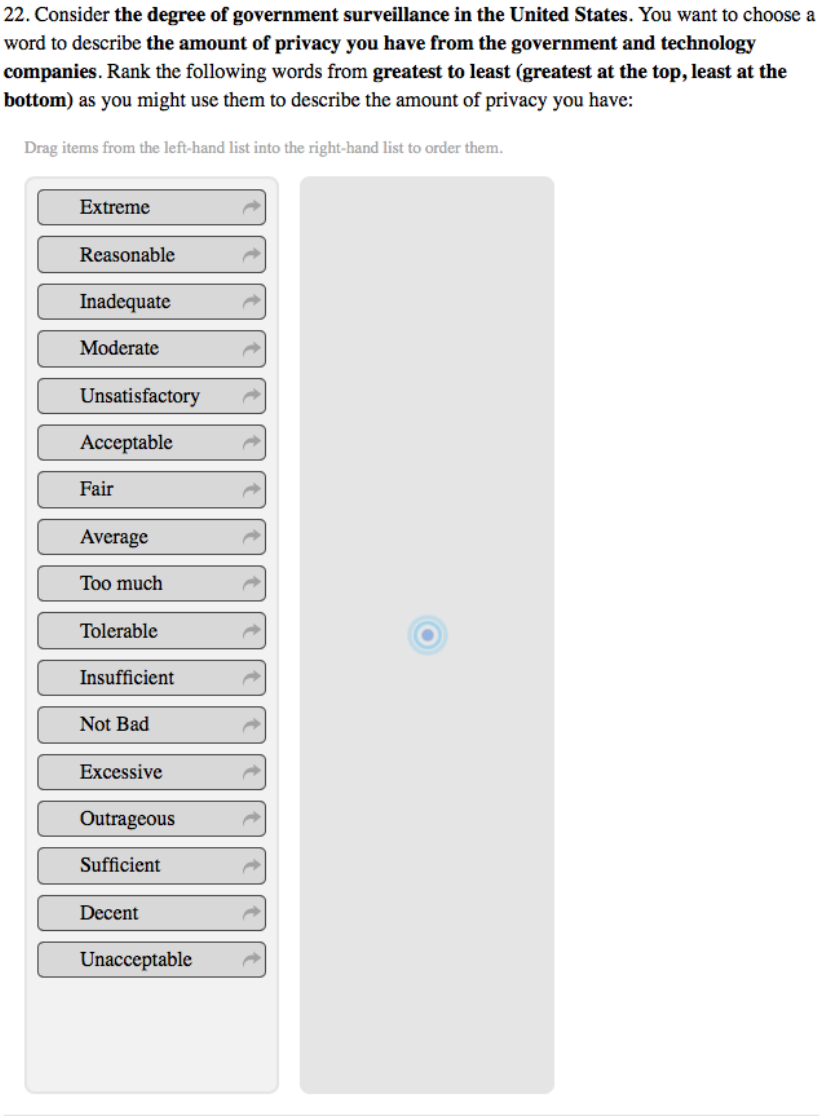


Figure 1: Word ranking task shown to participants

## 2.3 English Language Proficiency Screening

English proficiency of a participant is necessary to yield accurate results as a participant need to have the language capabilities to distinguish meanings among the list of words that contain synonyms. One could limit the AMT participation to people located in the U.S., but this is unreliable, because U.S. residency does not guarantee English proficiency, and there are ways to fake locations. Hence, we tested participants for English language reading proficiency using a subset of the Nelson-Denny reading comprehension test. The Nelson-Denny reading comprehension test is a standardized reading comprehension test used in the US and other countries to measure reading levels for students [2].

The complete Nelson-Denny test consists of 80 English vocabulary questions, followed by 38 comprehension questions based on six half-page essays and one full-page essay, that needs to be completed in 35-minute time period. Since using the full test could increase fatigue, we sought to identify a scored subset of the complete test that correlates with the complete scored test. We ran the complete test on 402 AMT participants, who were paid $6 to participate, and were allowed 45 minutes for completion. Mean score values for participants were 70 out of 80 (88%) for vocabulary and 31 out of 38 for reading comprehension (82%).

We analyzed the results using Multiple Correspondence Analysis (MCA) [Ler10] to select questions that are mostly representative of the variance in the sample. We selected the top 15 vocabulary questions and chose one half-page essay that contains the most representative questions, and used this subset as our screening test for the word-ranking study. Since these questions are still used in standardized tests, we will not share them in this report, but they can be made available for researchers for use upon request. The final subset of the Nelson-DennyâĂŹs test was used for screening participants. Participants from Mechanical Turk are only allowed to proceed to the word-ranking survey, if they pass the English language reading proficiency test in 15 minutes with a score of 80% or above in every section: vocabulary and reading comprehension.

# 3 Results of the Word-Ranking Study

We collected data from 205 participants who passed the English proficiency assessment with a score or 80% or more. Table 1 shows the mean rankings for each scenario along with the mean values associated with these rankings.

Results in Table 1 shows similarities in rankings among skewed scenarios. Note how the words outrageous, extreme and excessive rank in the bus and car scenarios that are intended to be skewed towards excessive, and how that changes for the meal size and privacy scenarios. In all the four scenarios, *excessive* ranked way greater than average; and except in the time waiting for bus scenario, *inadequate* ranked lower than average. This result makes us confident using *excessive* and *inadequate* as anchor points in semantic scales in surveys with average being the middle anchor point. In addition, we are especially interested in the results of the meals and privacy scenarios, because we intend to use the labels:*excessive* and *inadequate* in security-related scenarios which is also skewed towards inadequate.

Table 1: Mean word rankings for each scenario

| **Waiting for a bus** | | **Walking to car in parking** | | **Restaurant meal portions** | | **Privacy measures against surveillance** | |
|---|---|---|---|---|---|---|---|
| Word Rank | Mean | Word Rank | Mean | Word Rank | Mean | Word Rank | Mean |
| Outrageous | 2.82 | Outrageous | 3.14 | Extreme | 3.69 | Extreme | 5.77 |
| Extreme | 3.47 | Extreme | 3.21 | Excessive | 3.86 | Excessive | 5.91 |
| Excessive | 3.90 | Excessive | 4.06 | Outrageous | 4.40 | Outrageous | 6.00 |
| Too much | 5.24 | Too much | 5.63 | Too much | 4.56 | Too much | 6.37 |
| Unacceptable | 6.21 | Unacceptable | 6.98 | Sufficient | 7.91 | Moderate | 8.29 |
| Unsatisfactory | 7.40 | Unsatisfactory | 8.31 | Moderate | 8.04 | Reasonable | 8.92 |
| Inadequate | 10.94 | Moderate | 9.47 | Reasonable | 8.06 | Sufficient | 8.94 |
| Moderate | 10.03 | Tolerable | 9.88 | Acceptable | 8.14 | Average | 9.05 |
| Average | 10.35 | Average | 10.55 | Average | 8.32 | Decent | 9.05 |
| Tolerable | 10.35 | Reasonable | 10.69 | Decent | 8.36 | Acceptable | 9.21 |
| Insufficient | 11.46 | Inadequate | 11.07 | Fair | 9.55 | Fair | 9.55 |
| Sufficient | 11.78 | Decent | 11.31 | Tolerable | 10.54 | Tolerable | 9.95 |
| Acceptable | 11.79 | Acceptable | 11.32 | Not bad | 10.85 | Unsatisfactory | 10.66 |
| Reasonable | 11.51 | Sufficient | 11.40 | Inadequate | 14.03 | Unacceptabl | 10.77 |
| Fair | 11.77 | Fair | 11.46 | Unsatisfactory | 14.078 | Not bad | 11.00 |
| Decent | 11.71 | Insufficient | 11.49 | Insufficient | 14.082 | Inadequate | 11.53 |
| Not bad | 12.08 | Not bad | 12.28 | Unacceptable | 14.51 | Insufficient | 11.86 |

*Participants rated words from greatest at the top and lowest at the bottom; values from 1-17
with 1: greatest, 17: lowest

# 4 Conclusion

In this paper, we present an approach to evaluate linguistic labels before they are being used in applications. We provide a methodology that software engineers and computer scientist can follow when they design new applications.

# References


[1] D. Benton. Portion size: what we know and what we need to know. *Critical reviews in food science and nutrition*, 55(7):988–1004, 2015.

[2] J. I. Brown. The nelson-denny reading test. 1960.

[3] T. Dinev, P. Hart, and M. R. Mullen. Internet privacy concerns and beliefs about government surveillance–an empirical investigation. *The Journal of Strategic Information Systems*, 17(3):214–233, 2008.

[4] M. Furr. *Scale construction and psychometrics for social and personality psychology*. SAGE Publications Ltd, 2011.

[5] J. M. Mendel. Uncertain rule-based fuzzy logic system: introduction and new directions. 2001.

[6] J. M. Mendel and D. Wu. *Perceptual computing: aiding people in making subjective judgments*, volume 13. John Wiley & Sons, 2010.

[7] C. I. Mosier. A critical examination of the concepts of face validity. *Educational and Psychological Measurement*, 1947.