Presentation Title

Subtitle or Tagline

May 20, 2025

as for ANEW. The NRC VAD lexicon v1.0 (Mohammad, 2018a) is the largest manually created VAD lexicon (in any language), and the only one that was created via comparative annotations (instead of rating scales). It has entries for about 20,000 English words.

Our Work: In this paper, we describe how we substantially add to the NRC VAD lexicon v1.0 by obtained human ratings of valence, arousal, and dominance for about 25,000 additional English words and more than 10,000 common multiword expressions (MWEs). The scores are fine-grained real-valued numbers from -1 (lowest V, A, or D) to 1 (highest V, A, or D). We show that the annotations lead to reliable VAD score (split-half reliability scores of $r = 0.99$ for valence, $r = 0.98$ for arousal, and $r = 0.96$ for dominance.) We will refer to this lexicon with about 55k entries as the *NRC Valence, Arousal, and Dominance (VAD) Lexicon v2*.

All of the annotation tasks described in this paper were approved by our institution's review board, which examined the methods to ensure that they were ethical. Special attention was paid to obtaining informed consent and protecting participant anonymity. The NRC VAD Lexicon v2 is made freely available for research through our project webpage.[2]

2 Related Work

Primary Dimensions of Meaning and Affect: Osgood et al. (1957) asked human participants to rate words along dimensions of opposites such as *heavy–light, good–bad, strong–weak,* etc. Factor analysis of these judgments revealed that the three most prominent dimensions of meaning are evaluation (*good–bad*), potency (*strong–weak*), and activity (*active–passive*). Russell (1980, 2003) showed through similar analyses of emotion words that the three primary independent dimensions of emotions are valence or pleasure (positiveness–negativeness/pleasure–displeasure), arousal (active–passive), and dominance (dominant–submissive). They argue that individual emotions such as joy, anger, and fear are points in a three-dimensional space of valence, arousal, and dominance. It is worth noting that even though the names given by Osgood et al. and Russel et al. are different, they describe similar dimensions (Bakker et al., 2014).

Primary Dimensions of Social Cognition: In a similar vein, Social Psychology research has shown that *warmth (W)* (friendliness, trustworthiness, and sociability) and *competence (C)* (ability, power, dominance, and assertiveness) are core dimensions of social cognition and stereotypes (Fiske et al., 2002; Bodenhausen et al., 2012; Fiske, 2018; Abele et al., 2016; Koch et al., 2024). That is, human beings quickly and subconsciously judge (assess) other people, groups of people, and even their own selves along the dimensions of warmth and competence (likely because of evolutionary pressures (MacDonald, 1992; Eisenbruch and Krasnow, 2022). Assessing W and C was central to early human survival (e.g., to anticipate whether someone will help them build useful things or whether they might steal their resources). Note that even though the Social Psychology work prefers the term competence, it essentially refers to the same dimension that in Affective Science research is called dominance. Warmth is considered to be a primary component of valence, which in turn is evolutionarily central to the approach–avoid response, and so some researchers argue that the ability to assess of warmth emerges earlier than dominance/competence (Cuddy et al., 2007). This is the *primacy of valence* hypothesis.

The dimensions of W and C (or D) have been shown to have substantial implications on a wide variety of facets, including: interpersonal status (Swencionis et al., 2017), social class (Durante and Fiske, 2017), self-beliefs (Wojciszke et al., 2009), political perception (Fiske et al., 2014), child development (Roussos and Dunham, 2016), cultural analyses (Fiske and Durante, 2016), as well as professional and organizational outcomes, such as hiring, employee evaluation, and allocation of tasks and resources (Cuddy et al., 2011).

Existing Affect Lexicons: We already discussed several VAD lexicons in the Introduction such as Bradley and Lang (1999), Warriner et al. (2013), Moors et al. (2013), Võ et al. (2009), Redondo et al. (2007), and Mohammad (2018a). Other work has focused on creating sentiment lexicons, where words are marked for whether they denotate or connotate sentiment (but do not include fine valence scores or any information about arousal and dominance). Examples of such lexicons include the General Inquirer (Stone et al., 1966), MPQA (Wiebe et al., 2005), and the NRC Emotion Lexicon (Mohammad and Turney, 2013, 2010). The

[2] http://saifmohammad.com/WebPages/nrc-vad.html

NRC Emotion Lexicon also includes entries for whether a words is associated with any of the eight basic motions Plutchik (1980).

The NRC Emotion Intensity Lexicon (Mohammad, 2018b) has real-valued scores of intensity for the words in the NRC Emotion Lexicon associated with any of the eight emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. The NRC WorryWords Lexicon (Mohammad, 2024a) is a list of over 44,000 English words and real-valued scores indicating their associations with anxiety: from -3 (maximum calmness) to 3 (maximum anxiety).

Automatically Creating Affect Lexicons: There is growing work on automatically determining word–sentiment and word–emotion associations. These include Mohammad and Kiritchenko (2015); Mohammad (2012); Strapparava and Valitutti (2004); Yang et al. (2007); Yu et al. (2015); Staiano and Guerini (2014); Bandhakavi et al. (2021); Muhammad et al. (2023) to name just a few. These methods often assign a real-valued score representing the degree of association. The VAD Lexicon can be used to evaluate how accurately the automatic methods capture valence, arousal, and dominance.

3 Obtaining Human Ratings of Valence, Arousal, and Dominance

The keys steps in obtaining the new annotations were as follows:

1. selecting the terms to be annotated
2. developing the questionnaire
3. developing measures for quality control (QC)
4. annotating terms on a crowdsource platform
5. discarding data from outlier annotators (QC)
6. aggregating data from multiple annotators to determine the VAD association scores

We describe each of the steps below.

1. Term Selection. The NRC VAD Lexicon v1.0 already included a large number of common English words from many different sources:

- All terms in the NRC Emotion Lexicon (Mohammad and Turney, 2013, 2010). The NRC Emotion Lexicon has about 14,000 words that are annotated to indicate whether they are associated with any of the eight basic emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust (Plutchik, 1980) The NRC lexicon terms were in turn chosen by taking the content words that occur frequently in the Google n-gram corpus (Brants and Franz, 2006).
- All 4,206 terms in the positive and negative lists of the General Inquirer (Stone et al., 1966).
- All 1,061 terms listed in ANEW (Bradley and Lang, 1999).
- All 13,915 terms listed in the Warriner et al. (2013) lexicon.
- 520 words from the Roget's Thesaurus categories corresponding to the eight basic Plutchik emotions.[3]
- About 1000 high-frequency content terms, including emoticons, from the Hashtag Emotion Corpus (HEC) (Mohammad, 2012). All tweets in the The HEC include at least one of the eight basic emotion words as a hashtag word (e.g., *#anger, #sadness*, etc.).

The union of all of the above sets resulted in about 20k terms that were then annotated for valence, arousal, and dominance.

To add new terms we wanted again focus on common English terms not included in v1.0, and in addition, we wanted to include common phrases (multi-word expressions, light verb constructions, etc.). Finally, we wanted to include terms for which other linguistically interesting annotations already exists (such as concreteness and age of acquisition ratings). Therefore we included terms from the Prevalence dataset (Brysbaert et al., 2019). This dataset has prevalence scores (how widely a word is known by English speakers), determined directly by asking people, for 62,000 lemmas. We included a term if it was marked as known to at least 70% of the people who provided responses for the term. (From this set we removed terms that are common person names or city names.) This resulted in close to 25k unigrams. We also included $\sim$10.5k most common multi-word expressions from the Muraki et al. (2023) dataset. This dataset has concreteness ratings for about 62k English MWEs, as well as their frequencies in a subtitles corpus (Brysbaert et al., 2012).

2. VAD Questionnaires The questionnaires used to annotate the data were developed after several rounds of pilot annotations. Detailed directions,

[3] http://www.gutenberg.org/ebooks/10681

including notes directing respondents to consider predominant word sense (in case the word is ambiguous) and example questions (with suitable responses) were provided. (See Appendix.) The primary instruction and the questions presented to annotators are shown below.

VALENCE: Consider positive feelings (or positive sentiment) to be a broad category that includes:
positiveness / pleasure / goodness / happiness /
greatness / brilliance / superiority / health etc.
Consider negative feelings (or negative sentiment) to be a category that includes:
negativeness / displeasure /badness / unhappiness /
insignificance / terribleness / inferiority / sickness etc.
If you do not know the meaning of a word or are unsure, you can look it up in a dictionary (e.g., the Merriam Webster) or on the internet.
Quality Control
Some questions have pre-determined correct answers. If you mark these questions incorrectly, we will give you immediate feedback in a pop-up box. An occasional misanswer is okay. However, if the rate of misanswering is high (e.g., >20%), then all of one's HITs may be rejected.

Select the options that most English speakers will agree with.

Q1. <term> is often associated with:
3: very positive feelings
2: moderately positive feelings
1: slightly positive feelings
0: not associated with positive or negative feelings
-1: slightly negative feelings
-2: moderately negative feelings
-3: very negative feelings

AROUSAL: This task is about words and their association with activeness or arousal. Consider activeness or arousal to be a broad category that includes:
active, aroused, stimulated, excited, jittery, alert, etc.
Consider inactiveness or calmness to be a broad category that includes:
inactive, calm, unaroused, passive, relaxed, sluggish, etc.
This task is not about sentiment. (For example, something can be positive and inactive (such as flower), positive and active (such as exercise and party), negative and active (such as murderer), and negative and inactive (such as negligent).

DOMINANCE: This task is about words and their association with dominance, competence, control of situation, or powerfulness. Consider dominance, competence, control of situation, or powerfulness to be a broad category that includes:
dominant, competent, in control of the situation,
powerful, influential, important, autonomous, etc.
Consider submissiveness, incompetence, controlled by outside factors, or weakness to be a broad category that includes:
submissive, incompetent, not in control of the situation,
weak, influenced, cared-for, guided, etc.
This task is not about sentiment. (For example, something can be positive and weak (such as a flower petal) and something can be negative and strong (such as tyrant).

3. Quality Control Measures. About 2% of the data was annotated beforehand by the authors and interspersed with the rest. These questions are referred to as *gold* (aka *control*) questions. Half of the gold questions were used to provide immediate feedback to the annotator (in the form of a pop-up on the screen) in case they mark them incorrectly. We refer to these as *popup gold*. This helps prevent the situation where one annotates a large number of instances without realizing that they are doing so incorrectly. It is possible, that some annotators share answers to gold questions with each other (despite this being against the terms of annotation). Thus, the other half of the gold questions were also separately used to track how well an annotator was doing the task, but for these gold questions no popup was displayed in case of errors. We refer to these as *no-popup gold*.

4. Crowdsourcing. We setup the annotation tasks on the crowdsourcing platform, *Mechanical Turk*. In the task settings, we specified that we needed annotations from nine people for each word. We obtained annotations from native speakers of English residing around the world. Annotators were free to provide responses to as many terms as they wished. The annotation task was approved by our institution's review board.

Demographics: About 95% of the respondents who annotated the words live in USA. The rest were from India, United Kingdom, and Canada. The average age of the respondents was 34 years. Among those that disclosed their gender, about 53% were female, 47% were male.[4]

5. Filtering. If an annotator's accuracy on the gold questions (popup or non-popup) fell below 80%, then they were refused further annotation, and all of their annotations were discarded (despite being paid for). See Table 1 for summary statistics.

6. Aggregation. Every response was mapped to an integer from -3 (highly negative/inactive/submissive) to 3 (highly positive/active/dominant) as follows:

- highly positive/active/dominant: 3
- moderately positive/active/dominant: 2
- slightly positive/active/dominant: 1
- neither positive/active/dominant nor negative/inactive/submissive: 0
- slightly negative/inactive/submissive: -1
- moderately negative/inactive/submissive: -2
- highly negative/inactive/submissive: -3

The final score for each term is simply the av-

[4] Respondents were shown optional text boxes to disclose their demographic information as they choose; especially important for social constructs such as gender, in order to give agency to the respondents and to avoid binary language.

Version	#Words	#MWEs	#Total
v1.1 (2018)	19,839	132	19,971
v2.1 (2025)	44,928	10,073	55,001

Table 1: Number of terms in the NRC VAD Lexicon in version 1.1 and 2.1.

Version	Avg. #Annot.	SHR (ρ)	SHR (r)
valence	7.83	0.98	0.99
arousal	7.96	0.97	0.98
dominance	8.06	0.96	0.96

Table 2: Average number of annotations per word and split half reliability measured through both Spearman rank (ρ) and Pearson's (r) correlations. Scores in the 0.9s indicate high reliability.

erage score it received from each of the annotators. The scores were then linearly transformed to the interval: -1 (highest negativeness/inactivity/submissiveness) to 1 (highest positiveness/activity/dominance).

The terms and their VAD scores were added to the NRC VAD Lexicon v1 to create the NRC VAD Lexicon v2.

4 Reliability of the Annotations

A useful measure of quality is the reproducibility of the end result—repeated independent manual annotations from multiple respondents should result in similar scores. To assess this reproducibility, we calculate average *split-half reliability (SHR)* over 1000 trials. SHR is a common way to determine reliability of responses to generate scores on an ordinal scale (Weir, 2005). All annotations for an item are randomly split into two halves. Two separate sets of scores are aggregated, just as described in Section 3 (bullet 6), from the two halves. Then we determine how close the two sets of scores are (using a metric of correlation). This is repeated 1000 times and the correlations are averaged. The last two columns in Table 2 show the results (split half-reliabilities). Spearman rank and Pearson correlation scores of over 0.95 for V, A, and D indicate high reliability of the real-valued scores obtained from the annotations. (For reference, if the annotations were random, then repeat annotations would have led to an SHR of 0. Perfectly consistent repeated annotations lead to an SHR of 1. Also, similar past work on word–anxiety associations had SHR scores in the 0.8s (Mohammad, 2024b).)

5 Applications and Future Work

The large number of entries in the VAD Lexicon and the high reliability of the scores make it useful for a number of research projects and applications. We list a few below:

- Understanding valence, arousal, and dominance, and the underlying mechanisms; how VAD relate to our mind and body; how VAD change with age, socio-economic status, weather, green spaces, etc.
- Determining how VAD manifest in language; how language shapes our VAD; how culture shapes the language of VAD; etc.
- Tracking the degree of VAD towards targets of interest such as climate change, government policies, biological vectors, etc.
- Studying stereotypes and social cognition; using the dominance aka competence lexicon to study how competence assessment capabilities develop in children and to track perceptions of competence towards various targets of interest.
- Developing automatic systems for detecting VAD; To provide features for automatic sentiment or emotion detection systems. They can also be used to obtain sentiment-specific word embeddings and sentiment-specific sentence representations.
- To study the interplay between the categorical emotion model and the VAD model of affect. Much of the prior work has only explored one of the two models. The VAD lexicon can be used along with lists of words associated with emotions such as joy, sadness, fear, etc. to study the correlation of V, A, and D, with those emotions.
- To identify syllables that consistently tend to occur in words with high VAD scores. This has implications in understanding how some syllables and sounds have a tendency to occur in words referring to semantically related concepts. Identifying V, A, and D scores associated with syllables is also useful in generating names for literary characters and commercial products that have the desired affectual response.
- Studying VAD in story telling; its relationship with central elements of narratology such as conflict and resilience. To identify

high V, A, and D words in books and literature. To facilitate work of researchers in digital humanities. To facilitate work on literary analysis.

- As a source of gold (reference) scores, the entries in the VAD lexicon can be used in the evaluation of automatic methods of determining V, A, and D.
- The dataset is also of potential use to psychologists and evolutionary linguists interested in determining how evolution shaped the representation of the world around us, and why certain personality traits are associated with higher or lower shared understanding of valence, arousal, and dominance of words.

Apart from exploring the applications above, we are also interested in creating VAD lexicons for other languages, especially Chinese, Hindi, Arabic, Spanish, and German. We can then explore characteristics of valence, arousal, and dominance that are common across cultures.

6 Conclusions

We present here the NRC VAD Lexicon v2, which has human ratings of valence, arousal, and dominance for more 55,000 English terms. Compared to v1, it has entries for an additional ∼25k words. It also now includes for the first time entries for common multi-word expressions (∼10k). We provide a detailed description of how the terms were selected, the annotation process, and various measures for quality control. We show that the ratings are highly reliable (split-half reliability of over 0.95 for all three dimensions). The lexicon enables a wide variety of research in psychology, NLP, public health, digital humanities, and social sciences. It is made freely available for research through our project webpage.[5]

7 Limitations

The lexicon created is one of the largest that exist with wide coverage and a large number of annotators (thousands of people as opposed to just a handful). However, no lexicon can cover the full range of linguistic and cultural diversity in emotion expression. The lexicons are largely restricted to words that are most commonly used in Standard American English and they capture emotion associations as judged by American native speakers of English. Annotators on Mechanical Turk are not representative of the wider US population. However, obtaining annotations from a large number of annotators (as we do) makes the lexicon more resilient to individual biases and captures more diversity in beliefs. We see this work as a first step that paves the way for more work using responses from various other groups of people and in various other languages. See Mohammad (2023) for a detailed discussion of the limitations and best-practices in the use of emotion lexicons.

8 Ethics and Data Statement

The crowd-sourced task presented in this paper was approved by our Institutional Research Ethics Board. Our annotation process stored no information about annotator identity and as such there is no privacy risk to them. The individual words selected did not pose any risks beyond the risks of occasionally reading text on the internet. The annotators were free to do as many word annotations as they wished. The instructions included a brief description of the purpose of the task (Figures 1 through 9).

VAD assessments are complex, nuanced, and often instantaneous mental judgments. Additionally, each individual may use language to convey these assessments slightly differently. See Mohammad (2023) for a detailed discussion of ethical considerations when computationally analyzing emotions and VAD using emotion lexicons. We discuss below some of the notable considerations. (See Mohammad (2022) for a broader discussion of ethical considerations relevant to automatic emotion recognition.)

1. *Coverage:* We sampled a large number of English words from other lexical sources (which themselves sample from many sources). Yet, the words included do not cover all domains, genres, and people of different locations, socio-economic strata, etc. equally. It likely includes more of the vocabulary common in the United States with socio-economic and educational backgrounds that allow for technology access.
2. *Word Senses and Dominant Sense Priors:* Words when used in different senses and contexts may be associated with different degrees of VAD associations. The entries in in

[5] http://saifmohammad.com/WebPages/nrc-vad.html

the VAD Lexicon are indicative of the associations with the predominant senses of the words. This is usually not problematic because most words have a highly dominant main sense (which occurs much more frequently than the other senses). In specialized domains, some terms might have a different dominant sense than in general usage. Entries in the lexicon for such terms should be appropriately updated or removed. Further, any conclusions using the lexicon should be made based on relative change of associations using a large number of textual tokens. For example, if there is a marked increase in low-valence words from one period to the next, where each period has thousands of word tokens, then the impact of word sense ambiguity is minimal, and it is likely that some broader phenomenon is causing the marked increase in low-valence words. (See last two bullets.)

3. *Not Immutable:* The VAD scores do not indicate an inherent unchangeable attribute. The associations can change with time (e.g., the decrease in negativeness associated with *inter-race relationships* over the last 100 years), but the lexicon entries are fixed. They pertain to the time they are created. However, they can be updated with time.
4. *Socio-Cultural Biases:* The annotations for VAD capture various human biases. These biases may be systematically different for different socio-cultural groups. Our data was annotated by mostly US, Canadian, UK, and Indian English speakers, but even within these countries there are many diverse socio-cultural groups. Notably, crowd annotators on Amazon Mechanical Turk do not reflect populations at large. In the US for example, they tend to skew towards male, white, and younger people. However, compared to studies that involve just a handful of annotators, crowd annotations benefit from drawing on hundreds and thousands of annotators (such as this work).
5. *Inappropriate Biases:* Our biases impact how we view the world, and some of the biases of an individual may be inappropriate. For example, one may have race or gender-related biases that may percolate subtly into one's notions of VAD associated with words. Our dataset curation was careful to avoid words from problematic sources. We also ask people annotate terms based on what most English speakers think (as opposed to what they themselves think). This helps to some extent, but the lexicon may still capture some historical VAD associations with certain identity groups. This can be useful for some socio-cultural studies; but we also caution that VAD associations with identity groups be carefully contextualized to avoid false conclusions.
6. *Perceptions (not "right" or "correct" labels):* Our goal here was to identify common perceptions of WTS association. These are not meant to be "correct" or "right" answers, but rather what the majority of the annotators believe based on their intuitions of the English language.
7. *Avoid Essentialism:* When using the lexicon alone, it is more appropriate to make claims about VAD word usage rather than the VAD of the speakers. For example, *'the use of high-valence words in the context of the target group grew by 20%'* rather than *'valence in the target group grew by 20%'*. In certain contexts, and with additional information, the inferences from word usage can be used to make broader VAD claims.
8. *Avoid Over Claiming:* Inferences drawn from larger amounts of text are often more reliable than those drawn from small amounts of text. For example, *'the use of high-valence words grew by 20%'* is informative when determined from hundreds, thousands, tens of thousands, or more instances. Do not draw inferences about a single sentence or utterance from the VAD associations of its constituent words.
9. *Embrace Comparative Analyses:* Comparative analyses can be much more useful than stand-alone analyses. Often, VAD word counts and percentages on their own are not very useful. For example, *'the use of high-valence words grew by 20% when compared to [data from last year, data from a different person, etc.]'* is more useful than saying *'on average, 5 high-valence words were used in every 100 words'*.

We recommend careful reflection of ethical considerations relevant for the specific context of deployment when using the VAD lexicon.

References

Andrea E Abele, Nicole Hauke, Kim Peters, Eva Louvet, Aleksandra Szymkow, and Yanping Duan. 2016. Facets of the fundamental content dimensions: Agency with competence and assertiveness—communion with warmth and morality. *Frontiers in psychology*, 7:1810.

Iris Bakker, Theo van der Voordt, Peter Vink, and Jan de Boon. 2014. Pleasure, arousal, dominance: Mehrabian and russell revisited. *Current Psychology*, 33(3):405–421.

Anil Bandhakavi, Nirmalie Wiratunga, Stewart Massie, and Deepak P. 2021. Emotion-aware polarity lexicons for twitter sentiment analysis. *Expert systems*, 38(7):e12332.

Galen V Bodenhausen, Sonia K Kang, and Destiny Peery. 2012. Social categorization and the perception of social groups. *The Sage handbook of social cognition*, pages 311–329.

Margaret M Bradley and Peter J Lang. 1999. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical report, The Center for Research in Psychophysiology, University of Florida.

Thorsten Brants and Alex Franz. 2006. Web 1t 5-gram version 1. *Linguistic Data Consortium*.

Marc Brysbaert, Paweł Mandera, Samantha F McCormick, and Emmanuel Keuleers. 2019. Word prevalence norms for 62,000 english lemmas. *Behavior research methods*, 51:467–479.

Marc Brysbaert, Boris New, and Emmanuel Keuleers. 2012. Adding part-of-speech information to the subtlex-us word frequencies. *Behavior research methods*, 44:991–997.

Amy JC Cuddy, Susan T Fiske, and Peter Glick. 2007. The bias map: behaviors from intergroup affect and stereotypes. *Journal of personality and social psychology*, 92(4):631.

Amy JC Cuddy, Peter Glick, and Anna Beninger. 2011. The dynamics of warmth and competence judgments, and their outcomes in organizations. *Research in organizational behavior*, 31:73–98.

Federica Durante and Susan T Fiske. 2017. How social-class stereotypes maintain inequality. *Current opinion in psychology*, 18:43–48.

Adar B Eisenbruch and Max M Krasnow. 2022. Why warmth matters more than competence: A new evolutionary approach. *Perspectives on Psychological Science*, 17(6):1604–1623.

Susan Fiske, Amy Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82:878–902.

Susan T Fiske. 2018. Stereotype content: Warmth and competence endure. *Current directions in psychological science*, 27(2):67–73.

Susan T Fiske and Federica Durante. 2016. Stereotype content across cultures. *Handbook of advances in culture and psychology*, 6:209–258.

Susan T Fiske, Federica Durante, et al. 2014. Never trust a politician? collective distrust, relational accountability, and voter response. *Power, politics, and paranoia: Why people are suspicious of their leaders*, pages 91–105.

Alex Koch, Austin Smith, Susan T Fiske, Andrea E Abele, Naomi Ellemers, and Vincent Yzerbyt. 2024. Validating a brief measure of four facets of social evaluation. *Behavior Research Methods*, 56(8):8521–8539.

Kevin MacDonald. 1992. Warmth as a developmental construct: An evolutionary analysis. *Child development*, 63(4):753–773.

Saif Mohammad. 2012. #Emotional Tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 246–255, Montréal, Canada.

Saif Mohammad. 2018a. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Saif M. Mohammad. 2018b. Word affect intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan.

Saif M. Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.

Saif M. Mohammad. 2023. Best practices in the creation and use of emotion lexicons. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia. Association for Computational Linguistics.

Saif M. Mohammad. 2024a. Worrywords: Norms of anxiety association for 44,450 english words. In *Proceedings of The Annual Conference of the Empirical Methods on Natural Language Processing (EMNLP 2024, main)*, Miami, FL.

Saif M. Mohammad. 2024b. WorryWords: Norms of anxiety association for over 44k English words. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16261–16278, Miami, Florida, USA. Association for Computational Linguistics.

Saif M. Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

Saif M. Mohammad and Peter D. Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL-HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, LA, California.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Agnes Moors, Jan De Houwer, Dirk Hermans, Sabine Wanmaker, Kevin Van Schie, Anne-Laura Van Harmelen, Maarten De Schryver, Jeffrey De Winne, and Marc Brysbaert. 2013. Norms of valence, arousal, dominance, and age of acquisition for 4,300 dutch words. *Behavior research methods*, 45(1):169–177.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Alipio Jorge, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku, and Stephen Arthur. 2023. AfriSenti: A Twitter sentiment analysis benchmark for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981, Singapore. Association for Computational Linguistics.

Emiko J Muraki, Summer Abdalla, Marc Brysbaert, and Penny M Pexman. 2023. Concreteness ratings for 62,000 english multiword expressions. *Behavior research methods*, 55(5):2522–2531.

C.E. Osgood, Suci G., and P. Tannenbaum. 1957. *The measurement of meaning*. University of Illinois Press.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience*, 1(3):3–33.

Jaime Redondo, Isabel Fraga, Isabel Padrón, and Montserrat Comesaña. 2007. The spanish adaptation of anew (affective norms for english words). *Behavior research methods*, 39(3):600–605.

Gina Roussos and Yarrow Dunham. 2016. The development of stereotype content: The use of warmth and competence in assessing social groups. *Journal of Experimental Child Psychology*, 141:133–144.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145.

Jacopo Staiano and Marco Guerini. 2014. Depechemood: a lexicon for emotion analysis from crowd-annotated news. *arXiv preprint arXiv:1405.1605*.

Philip Stone, Dexter C. Dunphy, Marshall S. Smith, Daniel M. Ogilvie, and associates. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press.

Carlo Strapparava and Alessandro Valitutti. 2004. Wordnet-Affect: An affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*, pages 1083–1086, Lisbon, Portugal.

Jillian K Swencionis, Cydney H Dupree, and Susan T Fiske. 2017. Warmth-competence tradeoffs in impression management across race and social-class divides. *Journal of Social Issues*, 73(1):175–191.

Melissa LH Võ, Markus Conrad, Lars Kuchinke, Karolina Urton, Markus J Hofmann, and Arthur M Jacobs. 2009. The berlin affective word list reloaded (bawl-r). *Behavior research methods*, 41(2):534–538.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Joseph P Weir. 2005. Quantifying test-retest reliability using the intraclass correlation coefficient and the sem. *The Journal of Strength & Conditioning Research*, 19(1):231–240.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.

Bogdan Wojciszke, Andrea E Abele, and Wiesław Baryla. 2009. Two dimensions of interpersonal attitudes: Liking depends on communion, respect depends on agency. *European Journal of Social Psychology*, 39(6):973–990.

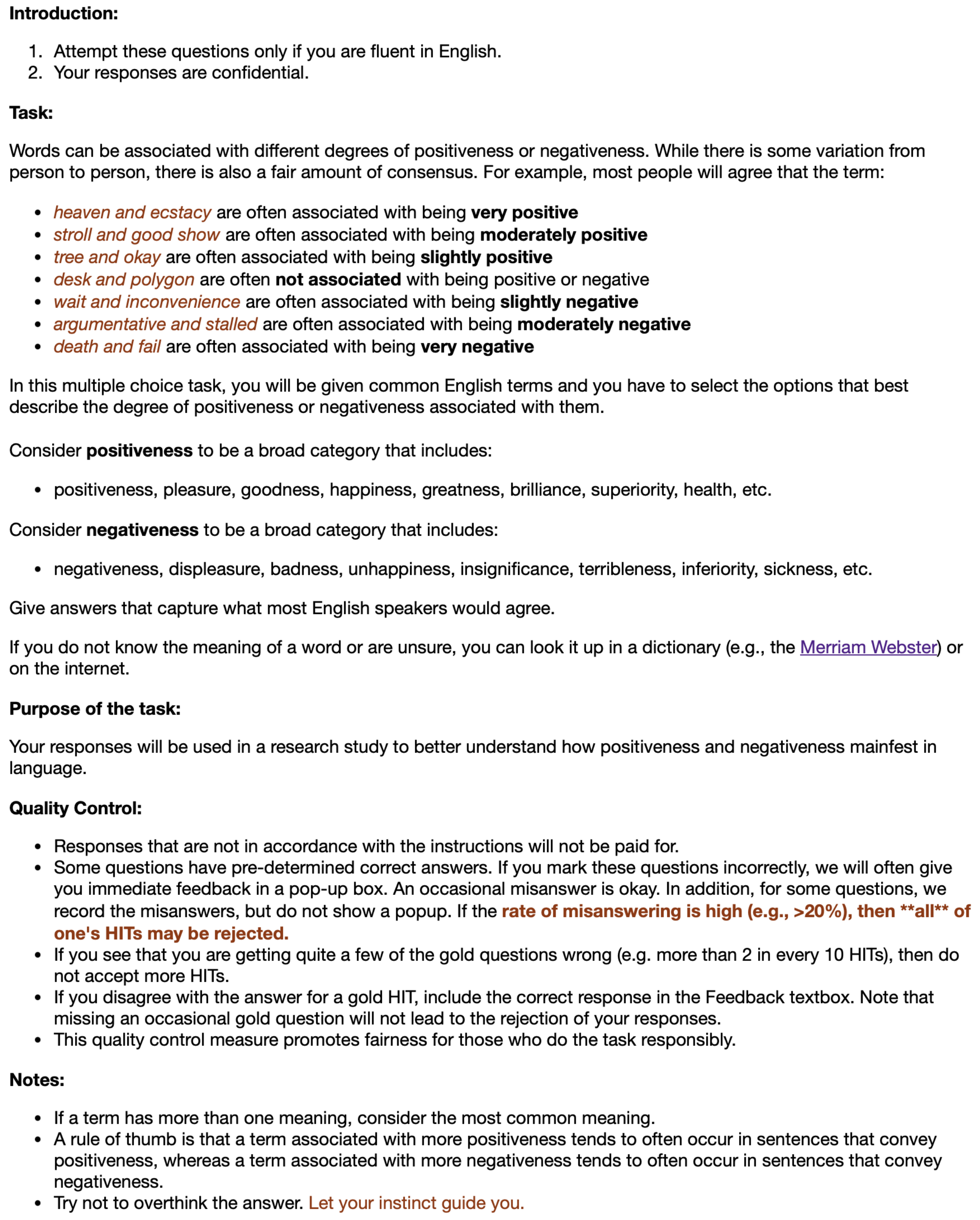
Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. 2007. Building emotion lexicon from weblog corpora. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 133–136.

Liang-Chih Yu, Jin Wang, K Robert Lai, and Xue-jie Zhang. 2015. Predicting valence-arousal ratings of words using a weighted graph method. In *Proceedings of the 53rd Annual Meeting of the Association* *for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 788–793.

A APPENDIX

A.1 AMT Questionnaires for Valence, Arousal, and Dominance

Screenshots of the detailed instructions, sample instance (question), and examples presented to the annotators are shown in Figures 1 through 9. Participants were informed that they may work on as many instances as they wish.



Introduction:

1. Attempt these questions only if you are fluent in English.
2. Your responses are confidential.

Task:

Words can be associated with different degrees of positiveness or negativeness. While there is some variation from person to person, there is also a fair amount of consensus. For example, most people will agree that the term:

- *heaven and ecstacy* are often associated with being **very positive**
- *stroll and good show* are often associated with being **moderately positive**
- *tree and okay* are often associated with being **slightly positive**
- *desk and polygon* are often **not associated** with being positive or negative
- *wait and inconvenience* are often associated with being **slightly negative**
- *argumentative and stalled* are often associated with being **moderately negative**
- *death and fail* are often associated with being **very negative**

In this multiple choice task, you will be given common English terms and you have to select the options that best describe the degree of positiveness or negativeness associated with them.

Consider **positiveness** to be a broad category that includes:

- positiveness, pleasure, goodness, happiness, greatness, brilliance, superiority, health, etc.

Consider **negativeness** to be a broad category that includes:

- negativeness, displeasure, badness, unhappiness, insignificance, terribleness, inferiority, sickness, etc.

Give answers that capture what most English speakers would agree.

If you do not know the meaning of a word or are unsure, you can look it up in a dictionary (e.g., the Merriam Webster) or on the internet.

Purpose of the task:

Your responses will be used in a research study to better understand how positiveness and negativeness mainfest in language.

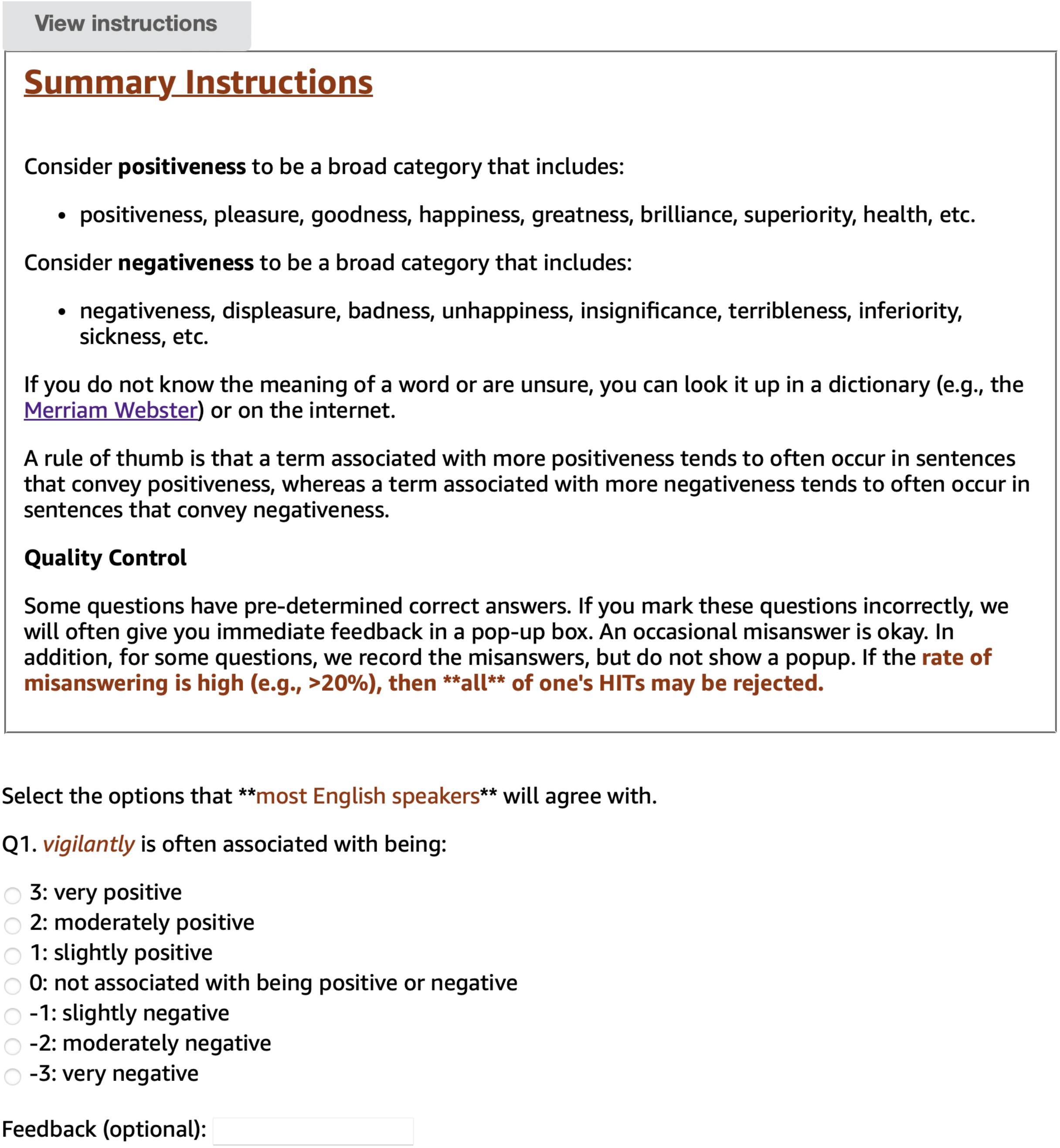
Quality Control:

- Responses that are not in accordance with the instructions will not be paid for.
- Some questions have pre-determined correct answers. If you mark these questions incorrectly, we will often give you immediate feedback in a pop-up box. An occasional misanswer is okay. In addition, for some questions, we record the misanswers, but do not show a popup. If the **rate of misanswering is high (e.g., >20%), then **all** of one's HITs may be rejected.**
- If you see that you are getting quite a few of the gold questions wrong (e.g. more than 2 in every 10 HITs), then do not accept more HITs.
- If you disagree with the answer for a gold HIT, include the correct response in the Feedback textbox. Note that missing an occasional gold question will not lead to the rejection of your responses.
- This quality control measure promotes fairness for those who do the task responsibly.

Notes:

- If a term has more than one meaning, consider the most common meaning.
- A rule of thumb is that a term associated with more positiveness tends to often occur in sentences that convey positiveness, whereas a term associated with more negativeness tends to often occur in sentences that convey negativeness.
- Try not to overthink the answer. Let your instinct guide you.

Figure 1: Valence Questionnaire: Detailed instructions.



View instructions

Summary Instructions

Consider **positiveness** to be a broad category that includes:

- positiveness, pleasure, goodness, happiness, greatness, brilliance, superiority, health, etc.

Consider **negativeness** to be a broad category that includes:

- negativeness, displeasure, badness, unhappiness, insignificance, terribleness, inferiority, sickness, etc.

If you do not know the meaning of a word or are unsure, you can look it up in a dictionary (e.g., the Merriam Webster) or on the internet.

A rule of thumb is that a term associated with more positiveness tends to often occur in sentences that convey positiveness, whereas a term associated with more negativeness tends to often occur in sentences that convey negativeness.

Quality Control

Some questions have pre-determined correct answers. If you mark these questions incorrectly, we will often give you immediate feedback in a pop-up box. An occasional misanswer is okay. In addition, for some questions, we record the misanswers, but do not show a popup. If the **rate of misanswering is high (e.g., >20%), then **all** of one's HITs may be rejected.**

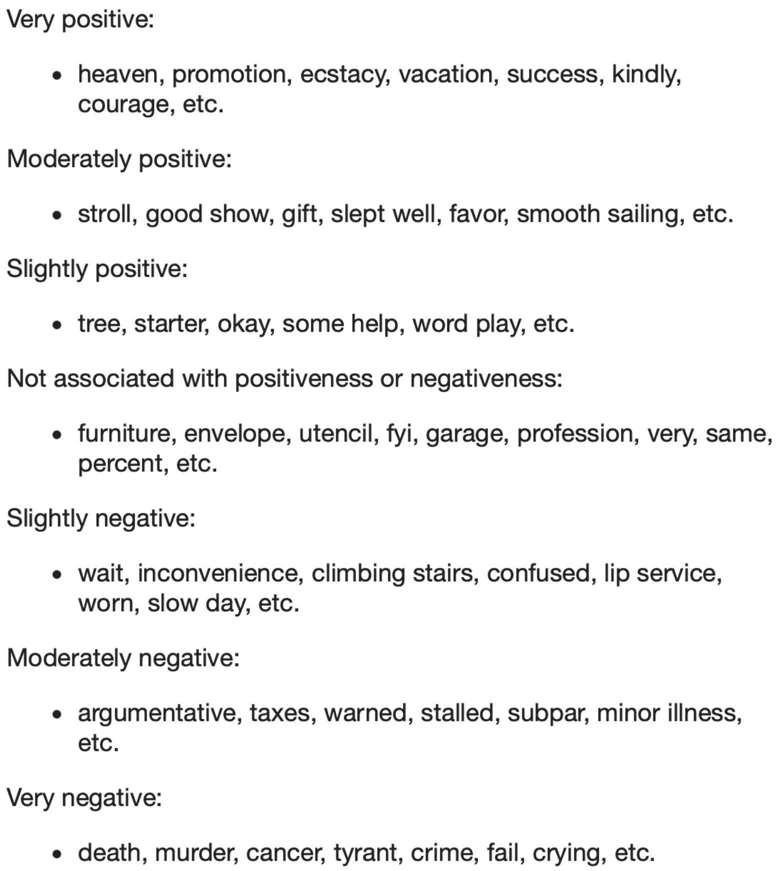
Select the options that **most English speakers** will agree with.

Q1. *vigilantly* is often associated with being:

○ 3: very positive
○ 2: moderately positive
○ 1: slightly positive
○ 0: not associated with being positive or negative
○ -1: slightly negative
○ -2: moderately negative
○ -3: very negative

Feedback (optional):

Figure 2: Valence Questionnaire: Sample question.



Very positive:

- heaven, promotion, ecstacy, vacation, success, kindly, courage, etc.

Moderately positive:

- stroll, good show, gift, slept well, favor, smooth sailing, etc.

Slightly positive:

- tree, starter, okay, some help, word play, etc.

Not associated with positiveness or negativeness:

- furniture, envelope, utencil, fyi, garage, profession, very, same, percent, etc.

Slightly negative:

- wait, inconvenience, climbing stairs, confused, lip service, worn, slow day, etc.

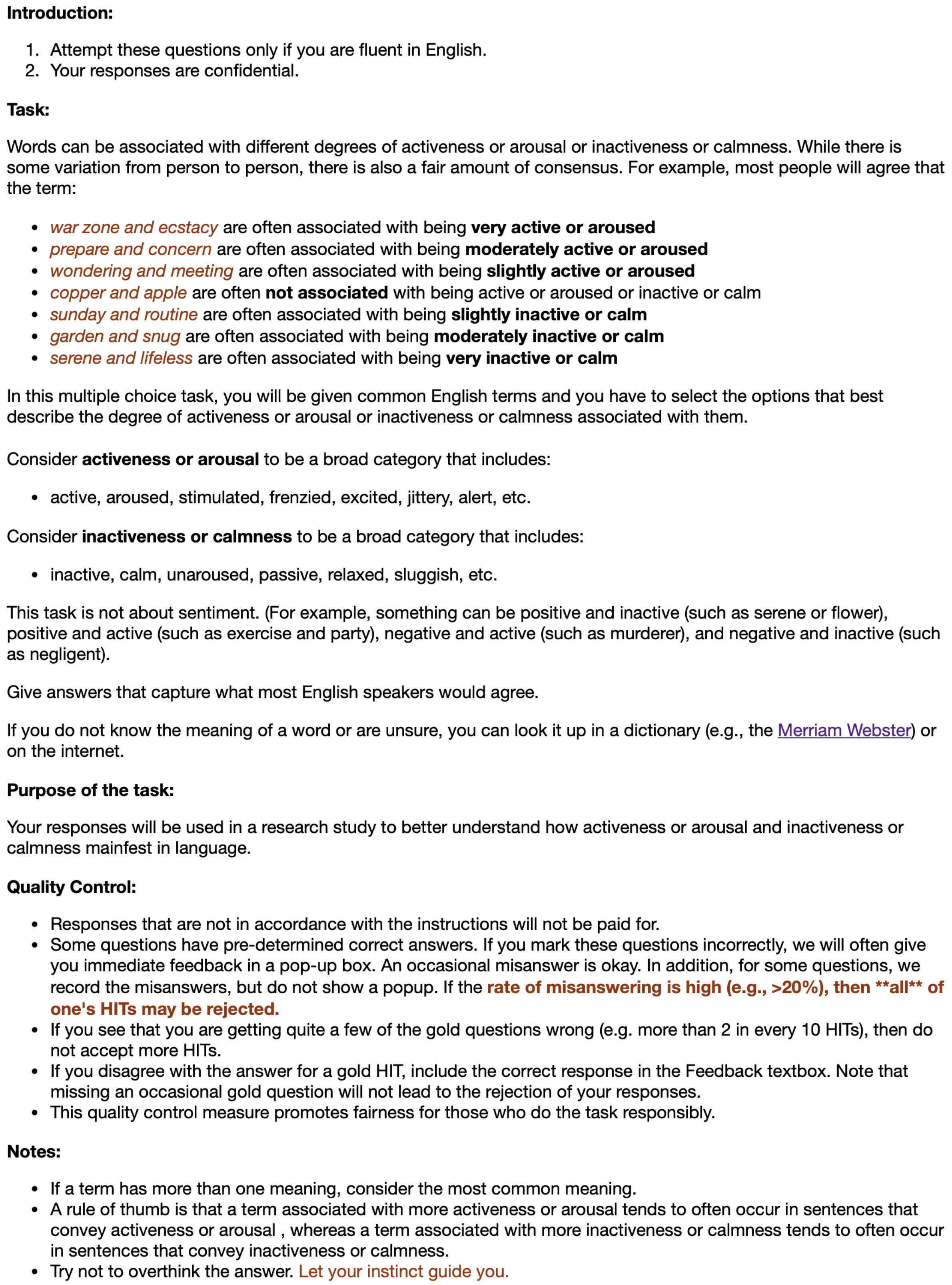
Moderately negative:

- argumentative, taxes, warned, stalled, subpar, minor illness, etc.

Very negative:

- death, murder, cancer, tyrant, crime, fail, crying, etc.

Figure 3: Valence Questionnaire: Examples.



Introduction:

1. Attempt these questions only if you are fluent in English.
2. Your responses are confidential.

Task:

Words can be associated with different degrees of activeness or arousal or inactiveness or calmness. While there is some variation from person to person, there is also a fair amount of consensus. For example, most people will agree that the term:

- *war zone and ecstacy* are often associated with being **very active or aroused**
- *prepare and concern* are often associated with being **moderately active or aroused**
- *wondering and meeting* are often associated with being **slightly active or aroused**
- *copper and apple* are often **not associated** with being active or aroused or inactive or calm
- *sunday and routine* are often associated with being **slightly inactive or calm**
- *garden and snug* are often associated with being **moderately inactive or calm**
- *serene and lifeless* are often associated with being **very inactive or calm**

In this multiple choice task, you will be given common English terms and you have to select the options that best describe the degree of activeness or arousal or inactiveness or calmness associated with them.

Consider **activeness or arousal** to be a broad category that includes:

- active, aroused, stimulated, frenzied, excited, jittery, alert, etc.

Consider **inactiveness or calmness** to be a broad category that includes:

- inactive, calm, unaroused, passive, relaxed, sluggish, etc.

This task is not about sentiment. (For example, something can be positive and inactive (such as serene or flower), positive and active (such as exercise and party), negative and active (such as murderer), and negative and inactive (such as negligent).

Give answers that capture what most English speakers would agree.

If you do not know the meaning of a word or are unsure, you can look it up in a dictionary (e.g., the Merriam Webster) or on the internet.

Purpose of the task:

Your responses will be used in a research study to better understand how activeness or arousal and inactiveness or calmness mainfest in language.

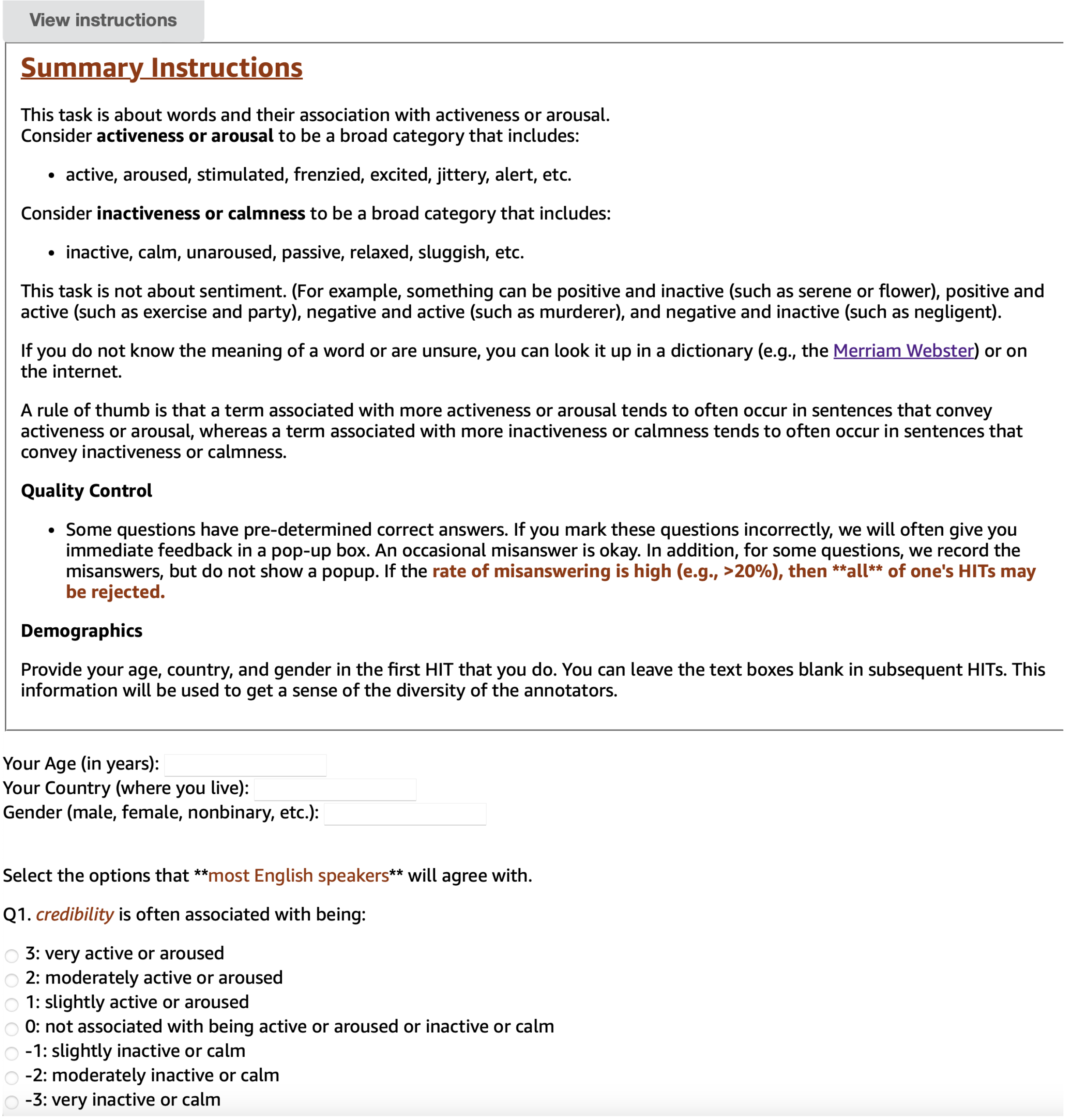
Quality Control:

- Responses that are not in accordance with the instructions will not be paid for.
- Some questions have pre-determined correct answers. If you mark these questions incorrectly, we will often give you immediate feedback in a pop-up box. An occasional misanswer is okay. In addition, for some questions, we record the misanswers, but do not show a popup. If the **rate of misanswering is high (e.g., >20%), then **all** of one's HITs may be rejected.**
- If you see that you are getting quite a few of the gold questions wrong (e.g. more than 2 in every 10 HITs), then do not accept more HITs.
- If you disagree with the answer for a gold HIT, include the correct response in the Feedback textbox. Note that missing an occasional gold question will not lead to the rejection of your responses.
- This quality control measure promotes fairness for those who do the task responsibly.

Notes:

- If a term has more than one meaning, consider the most common meaning.
- A rule of thumb is that a term associated with more activeness or arousal tends to often occur in sentences that convey activeness or arousal , whereas a term associated with more inactiveness or calmness tends to often occur in sentences that convey inactiveness or calmness.
- Try not to overthink the answer. Let your instinct guide you.

Figure 4: Arousal Questionnaire: Detailed instructions.



Summary Instructions

This task is about words and their association with activeness or arousal.
Consider **activeness or arousal** to be a broad category that includes:

- active, aroused, stimulated, frenzied, excited, jittery, alert, etc.

Consider **inactiveness or calmness** to be a broad category that includes:

- inactive, calm, unaroused, passive, relaxed, sluggish, etc.

This task is not about sentiment. (For example, something can be positive and inactive (such as serene or flower), positive and active (such as exercise and party), negative and active (such as murderer), and negative and inactive (such as negligent).

If you do not know the meaning of a word or are unsure, you can look it up in a dictionary (e.g., the Merriam Webster) or on the internet.

A rule of thumb is that a term associated with more activeness or arousal tends to often occur in sentences that convey activeness or arousal, whereas a term associated with more inactiveness or calmness tends to often occur in sentences that convey inactiveness or calmness.

Quality Control

- Some questions have pre-determined correct answers. If you mark these questions incorrectly, we will often give you immediate feedback in a pop-up box. An occasional misanswer is okay. In addition, for some questions, we record the misanswers, but do not show a popup. If the **rate of misanswering is high (e.g., >20%), then **all** of one's HITs may be rejected.**

Demographics

Provide your age, country, and gender in the first HIT that you do. You can leave the text boxes blank in subsequent HITs. This information will be used to get a sense of the diversity of the annotators.

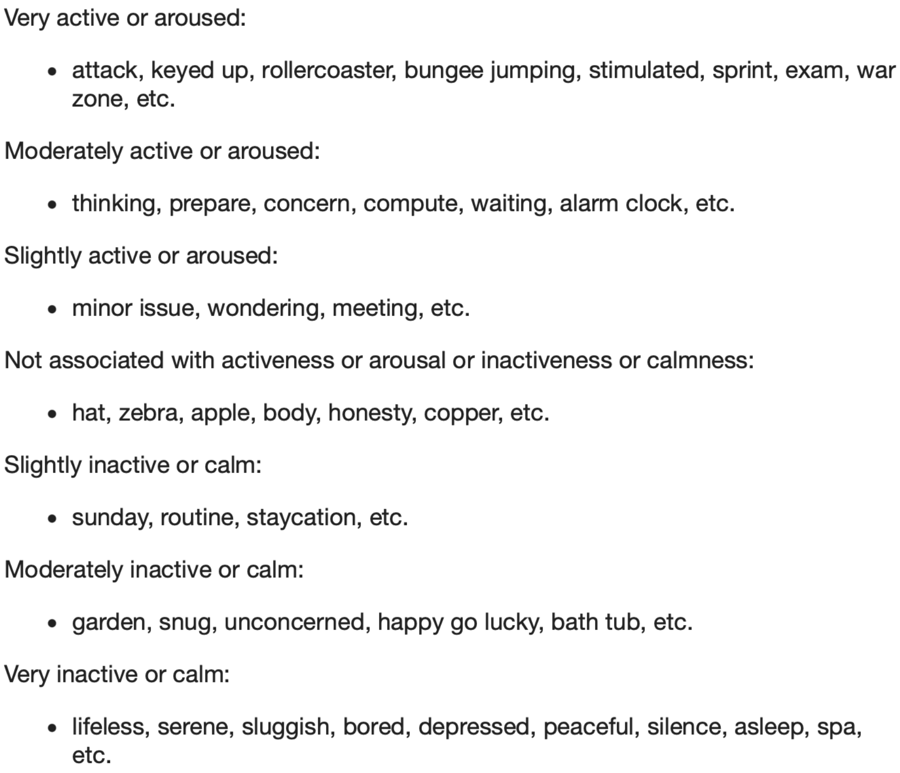
Your Age (in years):
Your Country (where you live):
Gender (male, female, nonbinary, etc.):

Select the options that **most English speakers** will agree with.

Q1. *credibility* is often associated with being:

○ 3: very active or aroused
○ 2: moderately active or aroused
○ 1: slightly active or aroused
○ 0: not associated with being active or aroused or inactive or calm
○ -1: slightly inactive or calm
○ -2: moderately inactive or calm
○ -3: very inactive or calm

Figure 5: Arousal Questionnaire: Sample question.



Very active or aroused:

- attack, keyed up, rollercoaster, bungee jumping, stimulated, sprint, exam, war zone, etc.

Moderately active or aroused:

- thinking, prepare, concern, compute, waiting, alarm clock, etc.

Slightly active or aroused:

- minor issue, wondering, meeting, etc.

Not associated with activeness or arousal or inactiveness or calmness:

- hat, zebra, apple, body, honesty, copper, etc.

Slightly inactive or calm:

- sunday, routine, staycation, etc.

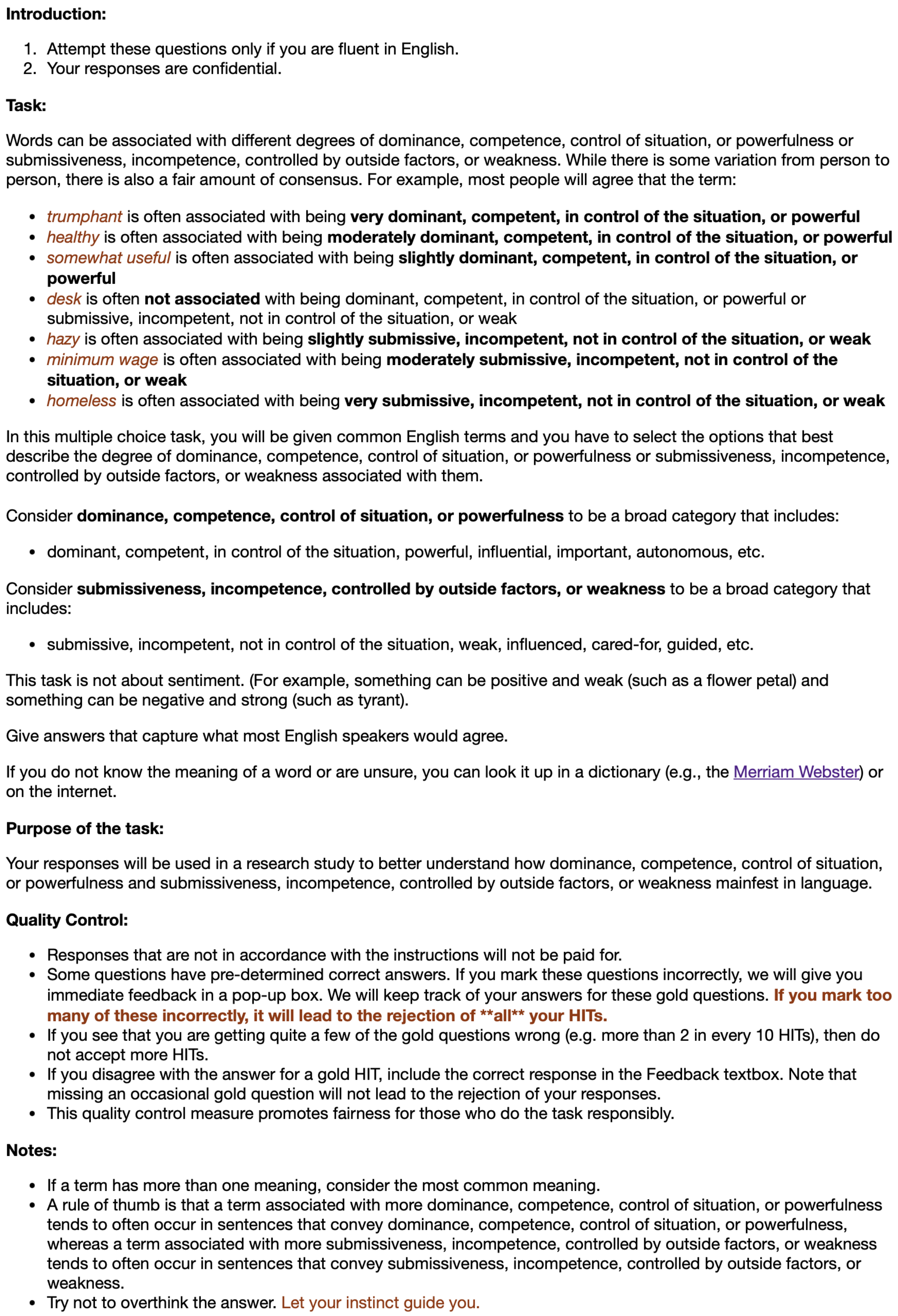
Moderately inactive or calm:

- garden, snug, unconcerned, happy go lucky, bath tub, etc.

Very inactive or calm:

- lifeless, serene, sluggish, bored, depressed, peaceful, silence, asleep, spa, etc.

Figure 6: Arousal Questionnaire: Examples.



Introduction:

1. Attempt these questions only if you are fluent in English.
2. Your responses are confidential.

Task:

Words can be associated with different degrees of dominance, competence, control of situation, or powerfulness or submissiveness, incompetence, controlled by outside factors, or weakness. While there is some variation from person to person, there is also a fair amount of consensus. For example, most people will agree that the term:

- *trumphant* is often associated with being **very dominant, competent, in control of the situation, or powerful**
- *healthy* is often associated with being **moderately dominant, competent, in control of the situation, or powerful**
- *somewhat useful* is often associated with being **slightly dominant, competent, in control of the situation, or powerful**
- *desk* is often **not associated** with being dominant, competent, in control of the situation, or powerful or submissive, incompetent, not in control of the situation, or weak
- *hazy* is often associated with being **slightly submissive, incompetent, not in control of the situation, or weak**
- *minimum wage* is often associated with being **moderately submissive, incompetent, not in control of the situation, or weak**
- *homeless* is often associated with being **very submissive, incompetent, not in control of the situation, or weak**

In this multiple choice task, you will be given common English terms and you have to select the options that best describe the degree of dominance, competence, control of situation, or powerfulness or submissiveness, incompetence, controlled by outside factors, or weakness associated with them.

Consider **dominance, competence, control of situation, or powerfulness** to be a broad category that includes:

- dominant, competent, in control of the situation, powerful, influential, important, autonomous, etc.

Consider **submissiveness, incompetence, controlled by outside factors, or weakness** to be a broad category that includes:

- submissive, incompetent, not in control of the situation, weak, influenced, cared-for, guided, etc.

This task is not about sentiment. (For example, something can be positive and weak (such as a flower petal) and something can be negative and strong (such as tyrant).

Give answers that capture what most English speakers would agree.

If you do not know the meaning of a word or are unsure, you can look it up in a dictionary (e.g., the Merriam Webster) or on the internet.

Purpose of the task:

Your responses will be used in a research study to better understand how dominance, competence, control of situation, or powerfulness and submissiveness, incompetence, controlled by outside factors, or weakness mainfest in language.

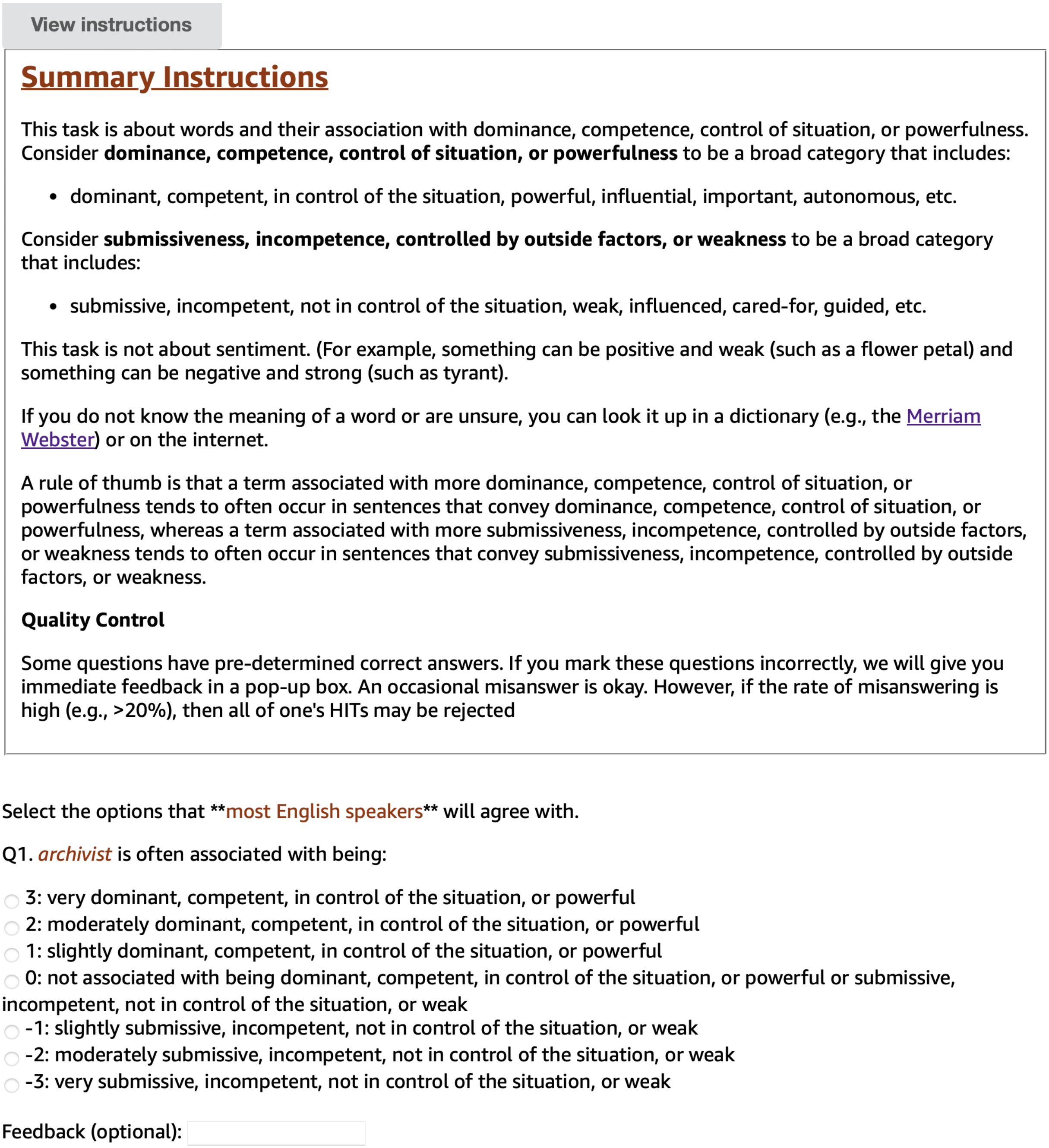
Quality Control:

- Responses that are not in accordance with the instructions will not be paid for.
- Some questions have pre-determined correct answers. If you mark these questions incorrectly, we will give you immediate feedback in a pop-up box. We will keep track of your answers for these gold questions. **If you mark too many of these incorrectly, it will lead to the rejection of **all** your HITs.**
- If you see that you are getting quite a few of the gold questions wrong (e.g. more than 2 in every 10 HITs), then do not accept more HITs.
- If you disagree with the answer for a gold HIT, include the correct response in the Feedback textbox. Note that missing an occasional gold question will not lead to the rejection of your responses.
- This quality control measure promotes fairness for those who do the task responsibly.

Notes:

- If a term has more than one meaning, consider the most common meaning.
- A rule of thumb is that a term associated with more dominance, competence, control of situation, or powerfulness tends to often occur in sentences that convey dominance, competence, control of situation, or powerfulness, whereas a term associated with more submissiveness, incompetence, controlled by outside factors, or weakness tends to often occur in sentences that convey submissiveness, incompetence, controlled by outside factors, or weakness.
- Try not to overthink the answer. Let your instinct guide you.

Figure 7: Dominance Questionnaire: Detailed instructions.



View instructions

Summary Instructions

This task is about words and their association with dominance, competence, control of situation, or powerfulness. Consider **dominance, competence, control of situation, or powerfulness** to be a broad category that includes:

- dominant, competent, in control of the situation, powerful, influential, important, autonomous, etc.

Consider **submissiveness, incompetence, controlled by outside factors, or weakness** to be a broad category that includes:

- submissive, incompetent, not in control of the situation, weak, influenced, cared-for, guided, etc.

This task is not about sentiment. (For example, something can be positive and weak (such as a flower petal) and something can be negative and strong (such as tyrant).

If you do not know the meaning of a word or are unsure, you can look it up in a dictionary (e.g., the Merriam Webster) or on the internet.

A rule of thumb is that a term associated with more dominance, competence, control of situation, or powerfulness tends to often occur in sentences that convey dominance, competence, control of situation, or powerfulness, whereas a term associated with more submissiveness, incompetence, controlled by outside factors, or weakness tends to often occur in sentences that convey submissiveness, incompetence, controlled by outside factors, or weakness.

Quality Control

Some questions have pre-determined correct answers. If you mark these questions incorrectly, we will give you immediate feedback in a pop-up box. An occasional misanswer is okay. However, if the rate of misanswering is high (e.g., >20%), then all of one's HITs may be rejected

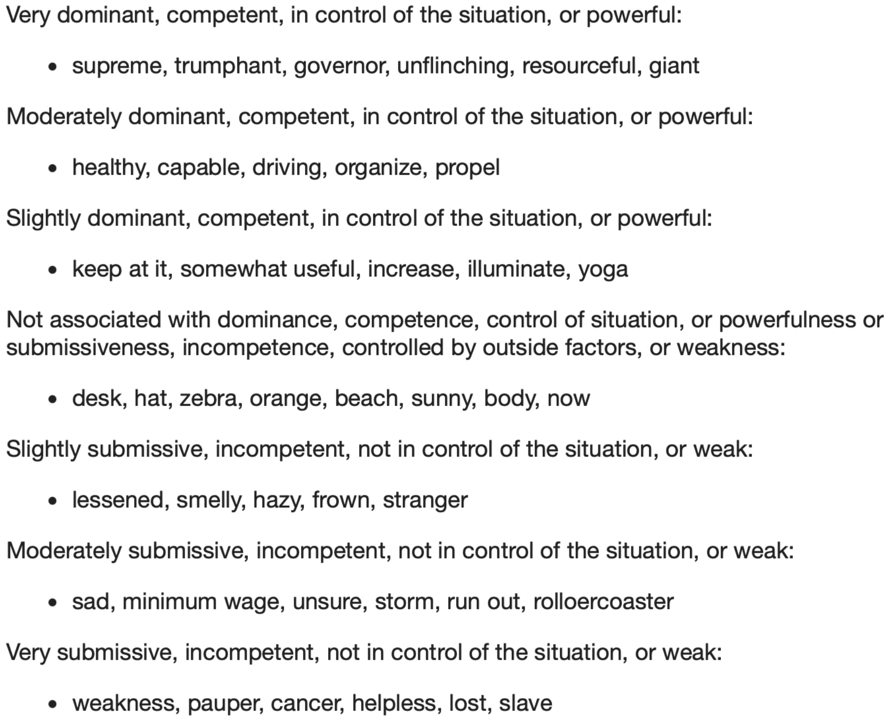
Select the options that **most English speakers** will agree with.

Q1. *archivist* is often associated with being:

○ 3: very dominant, competent, in control of the situation, or powerful
○ 2: moderately dominant, competent, in control of the situation, or powerful
○ 1: slightly dominant, competent, in control of the situation, or powerful
○ 0: not associated with being dominant, competent, in control of the situation, or powerful or submissive, incompetent, not in control of the situation, or weak
○ -1: slightly submissive, incompetent, not in control of the situation, or weak
○ -2: moderately submissive, incompetent, not in control of the situation, or weak
○ -3: very submissive, incompetent, not in control of the situation, or weak

Feedback (optional):

Figure 8: Dominance Questionnaire: Sample question.



Very dominant, competent, in control of the situation, or powerful:

- supreme, trumphant, governor, unflinching, resourceful, giant

Moderately dominant, competent, in control of the situation, or powerful:

- healthy, capable, driving, organize, propel

Slightly dominant, competent, in control of the situation, or powerful:

- keep at it, somewhat useful, increase, illuminate, yoga

Not associated with dominance, competence, control of situation, or powerfulness or submissiveness, incompetence, controlled by outside factors, or weakness:

- desk, hat, zebra, orange, beach, sunny, body, now

Slightly submissive, incompetent, not in control of the situation, or weak:

- lessened, smelly, hazy, frown, stranger

Moderately submissive, incompetent, not in control of the situation, or weak:

- sad, minimum wage, unsure, storm, run out, rolloercoaster

Very submissive, incompetent, not in control of the situation, or weak:

- weakness, pauper, cancer, helpless, lost, slave

Figure 9: Dominance Questionnaire: Examples.