Measuring trust: A text analysis approach to compare, contrast, and select trust questionnaires

Areen Alsaid[1*], Mengyao Li[2], Erin K. Chiou[3], John D. Lee[2]

[1]University of Michigan-Dearborn, Dearborn, MI, USA

[2]University of Wisconsin-Madison, Madison, WI, USA

[3]Arizona State University, Mesa, AZ, USA

*** Correspondence:**
Corresponding Author
alsaid@umich.edu

Keywords: Trust, Trust Assessment, Trust Measurement, Questionnaires, Text Analysis, Trust Layers

Abstract

Trust has emerged as a prevalent construct to describe relationships between people and between people and technology in myriad domains. Across disciplines, researchers have relied on many different questionnaires to measure trust. The degree to which these questionnaires differ has not been systematically explored. In this paper, we use a word-embedding text analysis technique to identify the differences and common themes across the most used trust questionnaires and provide guidelines for questionnaire selection. A mapping review was conducted to identify the existing trust questionnaires. In total, we included 46 trust questionnaires from three main domains (i.e., Automation, Humans, and E-commerce) with a total of 626 items measuring different trust layers (i.e., Dispositional, Learned, and Situational). Next, we encoded the words within each questionnaire using GloVe word embeddings and computed the embedding for each questionnaire item, and for each questionnaire. We reduced the dimensionality of the resulting dataset using UMAP to visualize these embeddings in scatterplots and implemented the visualization in a web app for interactive exploration of the questionnaires (link to app). At the word level, the semantic space serves to produce a lexicon of trust-related words. At the item and questionnaire level, the analysis provided recommendation on questionnaire selection based on the dispersion of questionnaires' items and at the domain and layer composition of each questionnaire. Along with the web app, the results help explore the semantic space of trust questionnaires and guide the questionnaire selection process. The results provide a novel means to compare and select trust questionnaires and to glean insights about trust from spoken dialog or written comments.

Trust Scales Explorer: https://areen.shinyapps.io/Trust_explorer/

1 Introduction

Trust has been studied in myriad contexts, from the internet to consumer products, healthcare, the military, and transportation. One challenge for advancing trust research is being able to measure trust precisely, and in a way that can generalize across contexts.

The study of trust in diverse contexts has resulted in multiple definitions: as a belief, an expectation or an attitude, an intention, and a behavior (Lee & See, 2004). While these definitions are conceptually distinct, they are also interrelated; beliefs are derived from an individual's past experiences, affective processing of beliefs governs attitudes, attitudes modulate intentions, and intentions are turned into behaviors (Ajzen & Fishbein, 1980). In the study of trust for systems design, trust is considered as a mediator between beliefs and behaviors, and hence, "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee & See, 2004).

Factors affecting trust, and general approaches to measuring trust, have been operationalized in multiple different ways. For example, an information processing perspective of trust considers purpose, process, and performance of an entity as the primary dimensions of information needed for a trustor to calibrate their trust in the entity (Chancey, Bliss, Yamani, & Handley, 2017; Lee & See, 2004). Related to purpose, process, and performance are the concepts of benevolence, integrity, and ability with more recent work focusing on how to measure specific dimensions of purpose and benevolence (Mayer, Davis, & Schoorman, 1995; Sheridan, 2019). A factor analysis of word associations has also shown that measuring general trust can fall on a continuum of trust-related and distrust-related words (Jian, Bisantz, Drury, & Llinas, 2000). A three-layer model of trust (Hoff & Bashir, 2014) suggests corresponding layers for trust measurement: dispositional, learned, and situational. The dispositional layer of trust refers to the individual tendency to trust others, the learned layer is based on the individual's prior experiences and interactions, and the situational layer is concerned with the level of trust given a specific situation and context. Because of the variation in how trust has been conceptualized and subsequently operationalized, there exists a multitude of trust measures of trust (Kohn, de Visser, Wiese, Lee, & Shaw, 2021).

Observable behaviors have been used as proxy measures of trust because they are often seen as more objective and less obtrusive than self-report measures. Behaviors are also often the main outcomes of interest when it comes to the study of trust, such as understanding what affects people's decisions to rely on the advice of a virtual real estate agent (Cassell & Bickmore, 2000), with trust being one important factor. Some behavioral measures used to study trust have included: compliance, reliance, eye gaze, voice, facial expression, and even pedal press intensity in automated vehicles (Lee, Liu, Domeyer, & DinparastDjadid, 2019; Meyer & Lee, 2013; Price, Lee, Dinparastdjadid, Toyoda, & Domeyer, 2017, Alsaid, 2020).

Behavioral measures of trust are useful for understanding trust as a socio-cognitive construct between agents interacting in real-time (Schilbach et al., 2013; Takayama, 2009) and can serve as inputs to models used to dynamically predict human behavior in specific task contexts (Domeyer, Venkatraman, Price, & Lee, 2018; Yang, Schemanske, & Searle, 2021). However, behavioral measures are often tied to a specific context or experimental setup and are considered indirect measures of trust because it is possible to engage in trust-related behaviors without actually involving trust. Therefore, when generalizing from laboratory studies, behavioral measures typically only *indicate* trust, contributing to potential misinterpretation or misapplication of a study's results, and risking construct validity that impedes advancing a theoretical basis of trust (Campbell & Fiske, 1959; Lee & See, 2004).

In contrast to behavioral measures, questionnaires (i.e., self-report) are often more direct measures of trust because trust is fundamentally an *attitude* and not a behavior. Therefore, asking a person about their attitude and closely associated factors, such as their beliefs and expectations, is important for understanding that person's trust. Although a person's reflective responses are also imperfect measures and not without limitations, questionnaires have the added advantage of being straightforward to administer, are typically rigorously developed based on trust theory, have established methods for validating empirically, and can more easily generalize across task contexts. Indeed, questionnaires have been widely used to measure trust. Yet, the literature indicates that several trust questionnaires have been developed for more specific task environments, perhaps to increase the sensitivity of the instrument. This has led to many trust questionnaires spanning multiple fields and contexts (Kohn, de Visser, Wiese, Lee, & Shaw, 2021).

The large pool of existing questionnaires presents a challenge to researchers in selecting the appropriate questionnaire. Questionnaire selection depends on several factors such as the application domain, the context, and the trust layer of interest. Questionnaire items can characterize different layers of trust such as dispositional, learned, and situational trust (Hoff & Bashir, 2014). A comparison between the questionnaires and their constituent items and words can guide the selection process. However, these relationships have not been systematically explored.

A recent paper described how nine questionnaires, measuring trust in automation specifically, related to one other based on a semantic network analysis of their constituent words (Jeong, Park, Park, Pham, & Lee, 2018). Using Latent Semantic Analysis (LSA) in combination with network analysis, the paper identified 14 highly central words that could be used to create an integrated scale. While promising, this paper focused on the similarity between keywords. Focusing on the words only might overlook the contextual information contained by a questionnaire item or a questionnaire as a whole. In this study, we investigate the similarity across the words, the questionnaire items, and the questionnaires.

Text analysis could be used to reveal connections between the many different trust questionnaires. These connections can be condensed and visualized in two-dimensional semantic spaces. The manifestation of these connections in the semantic space at different levels of analysis (i.e., words, items, and questionnaires) allows researchers to compare and select the words, items, or questionnaires that best support their research needs. Hence, text analysis thus provides one lens for considering the differences and similarities between various trust questionnaires

Accordingly, the present analysis is not aimed at developing a new scale, nor finding a single ideal one, as there is no single ideal questionnaire that works for all experiment and contexts (Kohn, de Visser, Wiese, Lee, & Shaw, 2021). However, it provides high-level comparison and guidance to researchers to choose the best-suited questionnaire for their research question.

1.1 A Primer on Text Analysis

In text analysis, words are often represented as embeddings. Embeddings are vectors of numbers that describe the location of a word in a high-dimensional semantic space relative to other words. For example, words like "cat" and "dog" would be closer to each other than "cat" and "mailbox". Words with similar meanings have similar vector representations and will thus be close to each other in the semantic space. The vector representation of words allows for mathematical operations that quantify the similarities of words and hence allows for advances in natural language processing applications like sentiment analysis and text autocompletion.

Methods that learn the vector representation of words are categorized into (i) global matrix factorization methods and (ii) local context window methods (Pennington, Socher, & Manning, 2014). The first method exploits statistical information contained by the words, such as Latent Semantic Analysis (LSA). The intuition is to extract relationships between the words in the corpus, assuming that words similar in meaning will appear in similar contexts (Landauer & Dumais, 2008). LSA relies on the frequency of word occurrence and ignores the context in which the words appear. It represents the text data in a corpus matrix that consists of word frequencies in each document. Each word occurrence in each document is counted, and the entire matrix is reduced using Singular Value Decomposition (SVD.) As a result, documents that share more words are considered similar, even if the similar words were used in a different context (e.g., the "bank" in "river bank" and "bank ATM" is considered equal (Hu, Zhang, & Zheng, 2016)). LSA produces semantic spaces that are high-level abstractions that are useful but lack context information.

The second method uses skip-gram models to capture the local context in which the word occurs. In Skip-gram models, a constant length window is moved along the corpus, and a neural network is trained to capture the co-occurrence of words in that entire window, and to predict context based on the central word (Altszyler, Sigman, Ribeiro, & Slezak, 2016). One example is a technique called word2Vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), which preserves the local context and provides a more precise description of the relationships between words compared to LSA and SVD. In word2Vec, embeddings are estimated by predicting words based on the words in the predefined window which enables the embeddings to capture relationships between words such that vector operations on the embeddings can complete word analogies in a meaningful fashion. In this paper, we use Global Vectors for Word Representation (GloVe) (Pennington et al., 2014). GloVe combines the benefits of global factorization and local context methods: it uses the statistical information contained by the words while also accounting for context. GloVe is trained on the non-zero elements of aggregated global word-to-word co-occurrence probability matrix and shows improved interpretability and accuracy compared to Word2Vec.

The vector representation of the words defines the position of each word in a high dimensional space, typically 100 to 500 dimensions. However, high-dimensional data is hard to visualize making it hard to identify what words similar to each other (Patel, 2016). Dimensionality reduction techniques reveal the underlying structure of the data. Principal component analysis (PCA) is a common dimensionality reduction technique that finds the linear combinations of the variables that capture the most variance in a dataset (Hubert, Rousseeuw, & Vanden Branden, 2005). t-Distributed Stochastic Neighbor Embedding (t-SNE) is another technique that accommodates non-linear relationships between the variables and more precisely captures the micro-structure of the data (Maaten & Hinton, 2008).

In this paper, we use a non-linear dimensionality reduction technique, called Uniform Manifold Approximation and Projection (UMAP), which captures non-linear relationships, like t-SNE, but in a more reproducible manner (McInnes, Healy, & Melville, 2018). An important difference between PCA and UMAP is that PCA preserves the global structure of the data whereas UMAP preserves the local structure. PCA ensures items that are different are distant from each other whereas UMAP ensures similar items are close to each other. This means that the inter-cluster relations can be more meaningful than in t-SNE. On the other hand, UMAP creates a low-dimensional representation where similar items are near each other, which is appropriate for this study. Since it is beneficial to understand both the inter-cluster and intra-cluster relationships for the trust questionnaires. For visualization purposes, we used a two-dimensional space to visualize how the questionnaires relate or diverge. However, the results of dimensionality reduction might not be directly comprehensible since it is highly non-linear after iterative topological data analysis (McInnes, Healy, Saul, & Großberger, 2018), nonetheless, they can reveal important relationships between the variables (Alsaid & Lee, 2022; Alsaid, Lee, Roberts, Barrigan, & Baldwin, 2018). For more details on text analysis and dimensionality reduction techniques, see Appendix A.

1.2 Research Objective

In this paper, we use word-embedding text analysis to understand different aspects of trust questionnaires and selecting the appropriate ones. First, we conduct a mapping literature review to gather existing questionnaires. Second, we apply text analysis techniques to quantify the relationships between the words used in the questionnaires, the questionnaire items within the questionnaires, and the overall questionnaires. These relationships were quantified using GloVe vector representations of the words. Third, we develop charts that quantify the composition of each questionnaire (i.e., application domain and trust layer composition) to guide researchers to select a questionnaire suited for the research task at hand. Finally, we generate a lexicon of the trust-related words that could be used to develop trust questionnaires and trust-focused sentiment analysis. The results are implemented in a web application that can help the researchers compare and contrast the different trust questionnaires and select the best fit for their research needs.

2 Method

2.1 Compiling and Labeling the Corpus

A mapping literature review (Grant & Booth, 2009) using Google Scholar was conducted using the keywords: “trust in automation, trust in humans, trust, e-commerce, assessment, scales”, their variants (e.g. “technology”, “robots”, “interpersonal trust”, “surveys”, “questionnaires”) and their combinations. Titles and abstracts were read to select those that developed or used rating-based trust measures. All *unique* developed questionnaires were included in the final selection and multiple questionnaires had overlapping items. A total of 80 articles were downloaded that met these inclusion criteria. Of these 80 articles, 46 questionnaires were extracted, with a total of 626 questionnaire items. After assessing the final selection of questionnaires, the questionnaires were categorized and labeled based on the domain for which they were developed (Chita-Tegmark, Law, Rabb, & Scheutz, 2021):

1. Automation: questionnaires developed for assessing trust in automation, including robots and technology more generally.

2. E-Commerce: questionnaires developed to assess consumers' trust in brands, trust in retailers' websites, and online shopping in general.

3. Human: questionnaires developed to assess interpersonal trust.

Because not all questionnaire items assessed the same layer of trust, the items within each questionnaire were also categorized and labeled according to the layer of trust that they measured, based on Hoff and Bashir's model of trust layers (2014), for its comprehensiveness:

1. Dispositional: measures a person's general tendency to trust, independent of context or a specific system. Dispositional trust arises from long-term biological and environmental influences.

2. Learned: measures a person's trust based on previous experiences based on a specific automated system

3. Situational: measures trust in a specific context or situation including both the external environment and the internal, context-dependent characteristics of the operator.

2.2 Data Cleaning

Once the corpus was compiled and labeled, the first step of our text analysis was to pre-process the data. We converted all words to lowercase, removed one-letter words, and punctuation. To focus the analysis on words relevant to trust assessment, we excluded stop words. Stop words refer to unimportant, uninformative, frequently used words such as pronouns, prepositions, and auxiliary verbs. Here, we used a list of stop words from the tidytext package, specifically, the Onix stop word lexicon. The Onix stop word list was moderately aggressive in removing words compared to other stop word lists. The Onix list includes words such as "become", "know", "fully", "great" and "interesting". Furthermore, we removed domain-specific words such as "shopping" and "product" that refer to trust in the retail industry. In addition, we removed words like "system" and "user" or "technology" and "consumer" because of their high frequency and limited relevance to making conceptual distinctions regarding trust. In ambiguous cases, e.g., cases in which a decision to include a word was unclear, we erred on the side of inclusion.

2.3 Data Analysis

Using the Wikipedia 2014 + Gigaword 5 pre-trained word vectors dataset provided on the GloVe website (Pennington et al., 2014), we calculated embeddings for each word, questionnaire item, and questionnaire. At the word level, we matched the words' vector embeddings with those in the pre-trained data. At the item level, we calculated the log odds ratios weighted by an uninformative Dirichlet prior for the words in each item. Using the log odds increases the weights for words that are common in a specific item, and relatively uncommon among all other items. This method gives greater weight to distinguishing words (Monroe, Colaresi, & Quinn, 2008). The log odds ratios were then used to create a weighted mean of the embeddings of the words that comprise each item. We used this same process to calculate an embedding for each questionnaire. For more details on the log odds ratio calculations see Appendix A.

To develop the trust lexicon, we calculated the log odds ratio of the words in the trust questionnaires, given the 5000 most common English words list from the wordfrequency website (Davies & Gardner, 2010a), and extracted the 20 most unique trust words. We calculated the cosine similarity distance between each of the 20 words and the GloVe word embeddings, similar to the approach of Fast, Chen, & Bernstein (2017), and for each word, we extracted the 20 closest words. Then we subjectively selected the 3-5 most related words based on our knowledge and judgment. For example, the word "aid" was used in the questionnaire in the context of "help". However, many of the close words were related to "funding" and "humanitarian" aid. Therefore, we kept the words that have a similar meaning to "aid" in the questionnaires such as "assistance" and "help". Typically, such lexicons are developed by labeling large corpora of text through crowdsourcing (Mohammad & Turney, 2013). Crowdsourcing is the practice of having a large number of people label data typically through the internet, a tedious and expensive process. However, the approach we use here is similar to the approach by Fast et al. (2017) which has shown to be highly accurate and efficient.

2.4 Software Tools

We used R statistical software (R Development Core Team, 2016) to create plots with the 'ggplot2' package (Wickham & Winston, 2019); for data cleaning, we used the 'tidyverse' (Wickham, 2016) and 'tidytext' (Silge & Robinson, 2016) packages, and for dimensionality reduction, we used the 'umap' package (Konopka, 2019).

3 Results

Table 1 shows the questionnaires included in the text analysis from our mapping review, the number of items in each questionnaire, the number of citations per article, and the labeled domain category.

3.1 Word-Level Analysis

Figure 1 shows the relevant trust words gleaned from the questionnaires' 626 items. The size of the word reflects its frequency; the bigger the word, the more often it occurs. The words are arrayed based on the UMAP dimensionality reduction of the word embeddings. The words near each other in this space are expected to have similar or complementary meanings. Some of the words create themes that directly map to different trust dimensions.

Table 1. Published Articles with Trust Questionnaires. Number of citations as of October 2021

Paper	Items	Citations	Category
Jian, Bisantz & Drury (2000)	12	1,139	Automation
Muir & Moray (1996)	9	1,014	Automation
Dzindolet, Peterson, Pomranky, Pierce & Beck (2003)	1	967	Automation
McKnight, Carter, Thatcher, & Clay (2011)	23	564	Automation
Madsen & Gregor (2000)	25	338	Automation
Singh, Molloy & Parasuraman (1993)	12	278	Automation
Yagoda & Gillan (2012)	36	155	Automation
Merritt (2011)	6	145	Automation
Schaefer (2013)	40	140	Automation
Charalambous, Fletcher & Webb (2016)	10	86	Automation
Chancey, Bliss, Yamani & Handley (2017)	15	61	Automation
Körber, Prasch & Bengler (2018)	1	58	Automation
Körber (2018)	19	54	Automation
Montague (2010)	28	52	Automation
Merritt, Unnerstall, Lee & Huber (2015)	10	51	Automation
Wiczorek & Manzey (2014)	1	50	Automation
Chien, Semnani-Azad, Lewis & Sycara (2014)	21	39	Automation
Goillau, Kelly, Boardman & Jeannot (2003)	8	21	Automation
Albert, Gribbons & Almadas (2009)	1	20	Automation
Walliser, deVisser & Shaw (2016)	1	18	Automation
Khalid, Shiung, Nooralishahi, Rasool, Helander, Kiong & Ai-vyrn (2016)	14	15	Automation
Malle & Ullman (2021)	20	12	Automation
Moeckli, Brown, Dow, Boyle, Schwarz & Xiong (2015)	12	11	Automation
Holthausen, Wintersberger, Walker & Riener (2020)	6	8	Automation
Salcedo, Ortiz, Lackey, Hudson & Taylor (2011)	20	5	Automation
Schneider, Jessup, Stokes, Rivers, Lohani & McCoy (2017)	6	5	Automation
Byrne & Marín (2018)	5	1	Automation
Gefen, Karahanna & Straub (2003)	7	9,222	E-Commerce
McKnight, Choudhury & Kacmar (2002)	3	5,981	E-Commerce
Bhattacherjee (2002)	7	1,835	E-Commerce
Delgado-Ballester, Munuera-Aleman & Yague-Guillen (2003)	8	1,202	E-Commerce
Sohaib & Kang (2015)	32	35	E-Commerce
Rotter (1967)	25	5,666	Human-Human
Rempel, Holmes & Zanna (1985)	26	4,947	Human-Human
Mayer & Davis (1999)	38	2,863	Human-Human
Yamagishi & Yamagishi (1994)	24	2,782	Human-Human
Gefen & Straub (2004)	6	2,296	Human-Human
Larzelere & Huston (1980)	8	1,989	Human-Human
Yamagishi (1986)	5	1,750	Human-Human
Johnson-George & Swap (1982)	25	1,630	Human-Human
Wrightsman (1964)	13	526	Human-Human
Miller & Mitamura (2003)	2	380	Human-Human
Evans & Revelle (2008)	21	332	Human-Human
Frazier, Johnson & Fainshmidt (2013)	4	86	Human-Human
Goto (1996)	3	74	Human-Human
Allen, Bergin & Pickar (2004)	7	33	Human-Human

For example, in the middle of the figure, there is a cluster that includes “believe”, “expect”, “advice”, and “decisions.” This area generally seems to be about anticipating future behavior and predictability. This contrasts with the bottom right cluster that focuses on fairness, and security and includes words like “fair”, “welfare”, “trust”, and “secure”. This is much more about how the behavior is valued. Also, the words in the upper right cluster include “experience”, “skills”, and “knowledge” which seem to characterize competence and ability. Finally, the upper left cluster includes words such as “dependable”, “reliable”, “competent”, “honest” which seem to characterize performance, and measure the integrity and reliability dimensions of trust. The upper left cluster also includes words like “cheat”, “honest”, and “sincere” which seem to characterize morality. Whether or not these provide a comprehensive account of trust is the topic for other papers (Chiou & Lee, 2021; Lee & See, 2004; Malle & Ullman, 2021), but it certainly gives us an idea about the current and most common state of how researchers are measuring trust perceptions. Generally, these dimensions of trust characterize trustworthiness in different objects, for example, integrity typically characterizes trust in humans whereas competence characterizes trust in automation (Malle & Ullman, 2021)

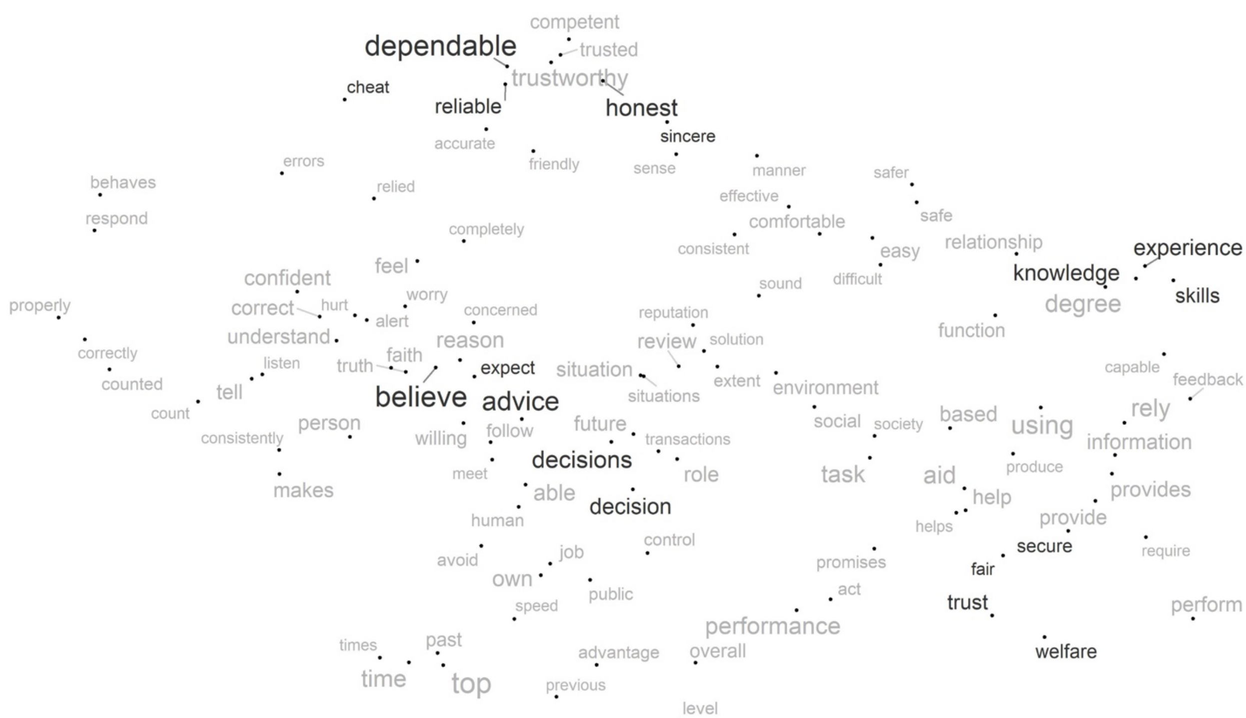

Figure 1. The UMAP two-dimensional representation of the questionnaire words. The words discussed in the manuscript as examples of specific themes are highlighted in black.

3.2 Item-Level Analysis

Figure 2 shows the UMAP representation of the items. The items from the most cited questionnaire in each domain are encircled and color-coded. Note that the figure shows *all* items of *all* questionnaires. Rectangles are also drawn around the items from the most cited questionnaires in each domain: the Rotter (1967) questionnaire was developed to measure human-human trust, the Gefen, Karahanna, & Straub (2003) questionnaire was developed to measure trust in online shopping, and the Muir & Moray (1996) questionnaire was designed to measure trust in automation. These different purposes are reflected in their placement in the semantic place.

Although very frequently cited, the questionnaire by Gefen et al. (2003) spans a small area, mainly because the items were similar and were developed to assess trust in an online vendor (i.e., "e-Commerce") based on past experiences. The questionnaire items were framed as, "Based on my experience with the online vendor in the past, I know it is…", with items assessing factors such as predictability and trustworthiness. On the other hand, the Rotter (1967) questionnaire has items that are widely spread across the semantic space. This is because it assesses dispositional trust (i.e., "Human-Human" trust) through a large set of questions related to views of the future, attitudes toward society, and hypothetical ethical scenarios. Muir & Moray's (1996) trust in automation questionnaire (i.e., "Automation") has some items close in the UMAP space to those of Rotter's interpersonal trust questionnaire; this is because it includes items that assess the reliability and dependability of automation and people. Interestingly, many questionnaires did not assess "trust" directly. Table 2 shows the number of items in each questionnaire that directly asked about trust.

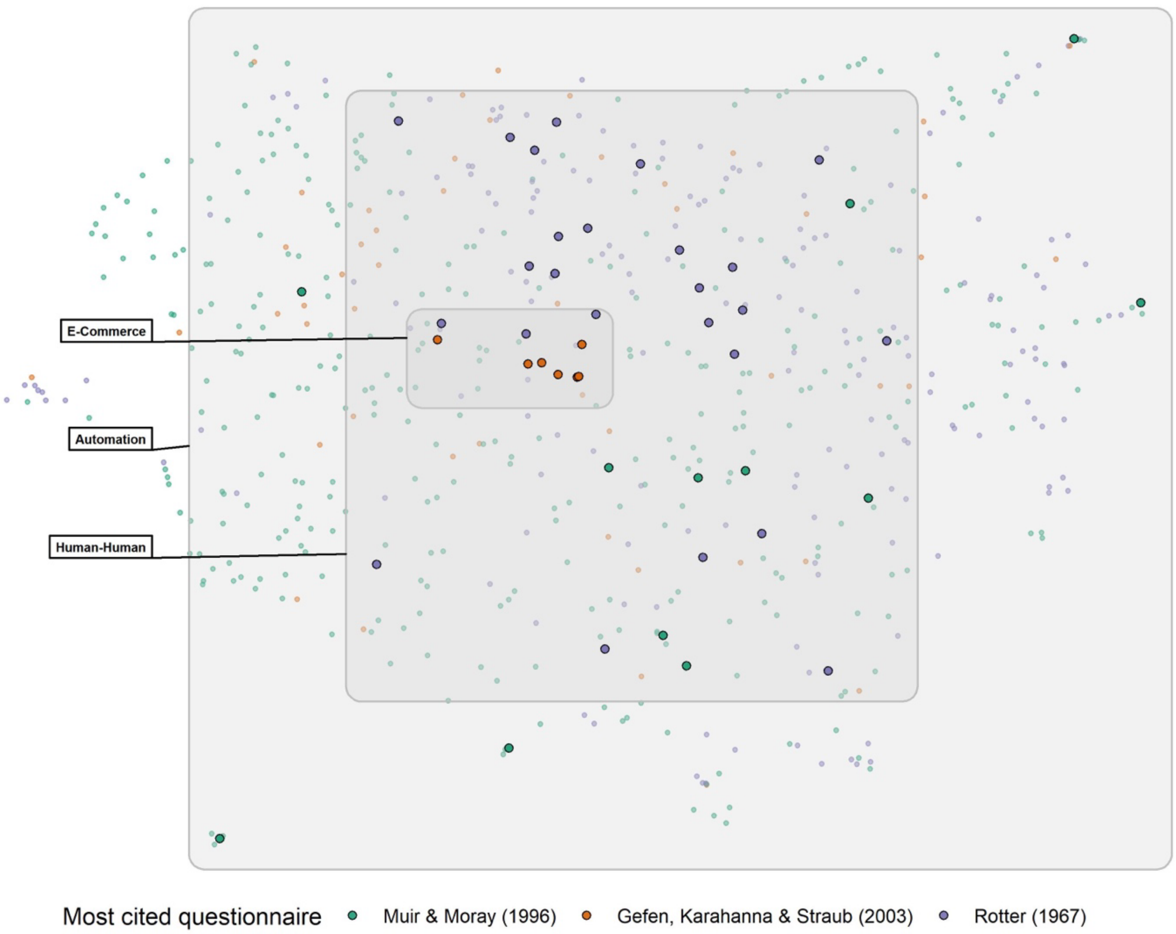

Figure 2. The UMAP two-dimensional representation of the embeddings of the questionnaire items. The most cited questionnaire in each category (Gefen, Karahanna & Straub, 2003 for e-commerce; Muir & Moray 1996 for automation; Rotter, 1967 for human-human). Green, orange, and purple represent items in automation, e-commerce, and human-human trust questionnaires, respectively.

Figure 2 showed that questionnaires vary in how their items spread across the semantic space. The questionnaire's spread is an indicator of the breadth of the questionnaire, and the different dimensions of trust it covers. Table 2 shows the spread value of each questionnaire. Spread was calculated as the average Euclidian distance between the questionnaire's items and the questionnaire centroid in the semantic space. The table also shows the number of items directly assessing trust (i.e., items that included the word trust specifically.)

Table 2. Questionnaires' spread measured as the mean Euclidian distance between questionnaires items in the semantic space

Paper	Category	Spread	Items including trust
Albert, Gribbons & Almadas (2009)	Automation	0	1
Dzindolet, Peterson, Pomranky, Pierce & Beck (2003)	Automation	0	0
Körber, Prasch & Bengler (2018)	Automation	0	2
Walliser, deVisser & Shaw (2016)	Automation	0	1
Wiczorek & Manzey (2014)	Automation	0	1
Chancey, Bliss, Yamani & Handley (2017)	Automation	3.2787649	2
Goillau, Kelly, Boardman & Jeannot (2003)	Automation	9.8178918	1
Merritt (2011)	Automation	11.0535503	5
Singh, Molloy & Parasuraman (1993)	Automation	13.3077262	0
Holthausen, Wintersberger, Walker & Riener (2020)	Automation	19.8532091	1
Merritt, Unnerstall, Lee & Huber (2015)	Automation	24.075077	0
Moeckli, Brown, Dow, Boyle, Schwarz & Xiong (2015)	Automation	26.7285164	5
Yagoda & Gillan (2012)	Automation	26.8700501	0
Khalid, Shiung, Nooralishahi, Rasool, Helander, Kiong & Ai-vyrn (2016)	Automation	26.9118744	3
Charalambous, Fletcher & Webb (2016)	Automation	27.3718082	0
Byrne & Marín (2018)	Automation	32.1872312	1
Salcedo, Ortiz, Lackey, Hudson & Taylor (2011)	Automation	35.4811105	0
Muir & Moray (1996)	Automation	40.9128248	3
Schneider, Jessup, Stokes, Rivers, Lohani & McCoy (2017)	Automation	44.2913665	2
Madsen & Gregor (2000)	Automation	45.7479433	0
Jian, Bisantz & Drury (2000)	Automation	52.9878988	1
Chien, Semnani-Azad, Lewis & Sycara (2014)	Automation	58.3499561	0
Malle & Ullman (2021)	Automation	63.6366837	0
Mcknight, Carter, Thatcher & Clay (2011)	Automation	64.6944447	0
Körber (2018)	Automation	72.247291	1
Montague (2010)	Automation	117.608309	2
Schaefer (2013)	Automation	141.482416	0
McKnight, Choudhury & Kacmar (2002)	E-Commerce	3.7473223	2
Gefen, Karahanna & Straub (2003)	E-Commerce	4.047058	1
Bhattacherjee (2002)	E-Commerce	14.6113045	1
Delgado-Ballester, Munuera-Aleman & Yague-Guillen (2003)	E-Commerce	29.7256945	1
Sohaib & Kang (2015)	E-Commerce	73.1215945	2
Goto (1996)	Human-Human	0.1458184	0
Miller & Mitamura (2003)	Human-Human	1.1019821	2
Frazier, Johnson & Fainshmidt (2013)	Human-Human	7.5817419	4
Yamagishi (1986)	Human-Human	19.049358	0
Gefen & Straub (2004)	Human-Human	21.7186439	2
Wrightsman (1964)	Human-Human	23.8513333	0
Evans & Revelle (2008)	Human-Human	30.5084731	0
Larzelere & Huston (1980)	Human-Human	31.4059937	2
Rempel, Holmes & Zanna (1985)	Human-Human	43.247883	1
Johnson-George & Swap (1982)	Human-Human	43.3757231	0
Allen, Bergin & Pickar (2004)	Human-Human	44.0729489	2
Rotter (1967)	Human-Human	51.903995	1
Yamagishi & Yamagishi (1994)	Human-Human	85.8964228	7
Mayer & Davis (1999)	Human-Human	122.881738	0

3.3 Questionnaire-Level Analysis

Figure 3 shows the UMAP semantic space representation of the questionnaires, which shows how the questionnaires relate to each other. The encircled areas highlight manually selected clusters. The upper part of the semantic space is dominated by trust in automation questionnaires, whereas the lower left bottom part mostly consists of human-human trust questionnaires. The E-Commerce questionnaires are spread across the entire space.

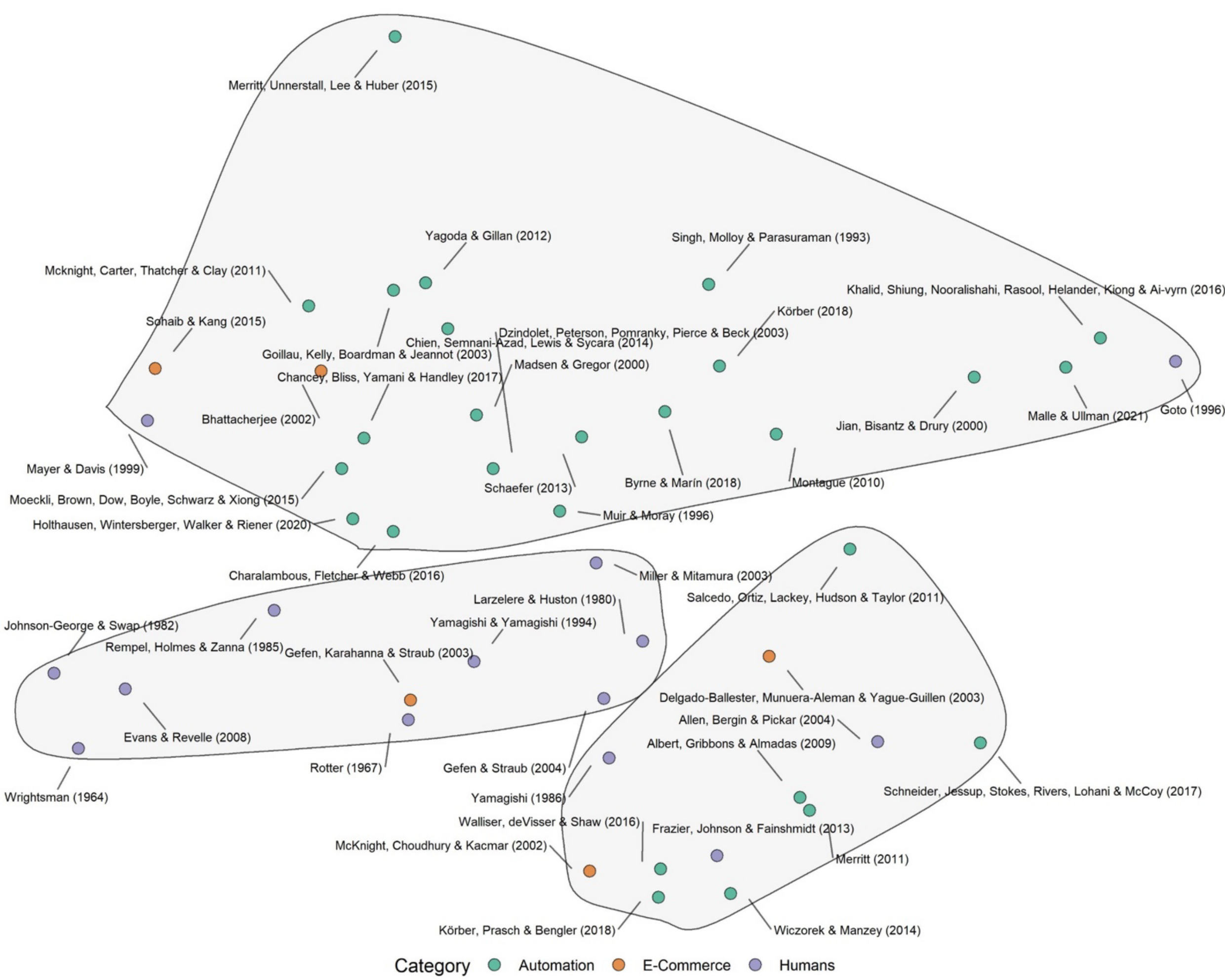

Figure 3.The UMAP two-dimensional representation of the questionnaire embeddings with domain category color-coded. Green, orange, and purple represent the automation, e-commerce, and human-human categories, respectively

Questionnaires close to each other share terms or similar terms that make them close in the UMAP space. The bottom left cluster consisted of questionnaires about trust in humans and had many questions related to peoples' behaviors such as "honesty" and "cheating" and how the person viewed "relationships" with and "personalities" of others. The questionnaires in the upper cluster mostly assessed efficiency, dependability, reliability, and safety in the specific domain that the questionnaire was developed. Finally, the questionnaires in the bottom right cluster commonly asked about the general tendency to trust (e.g., “I usually trust machines until there is a reason not to”.) The similarities and differences between these questionnaires can be further explored ls in the web app.

By calculating the log odds ratio of words in a specific category given all words used in all questionnaires, we identified the words that are most unique and distinguishing of trust across domains. Figure 4 shows the 10 highest frequencies of words in each domain (along the horizontal axis) and how unique these words are to each of the specific domains (along the vertical axis.) Some words had tied frequencies, and thus the figure shows more than 10 words in E-Commerce and Automation domains. The log odds ratio emphasizes words that are common in a specific category and relatively uncommon in others. For more details on the log odds ratio calculations see Appendix A.

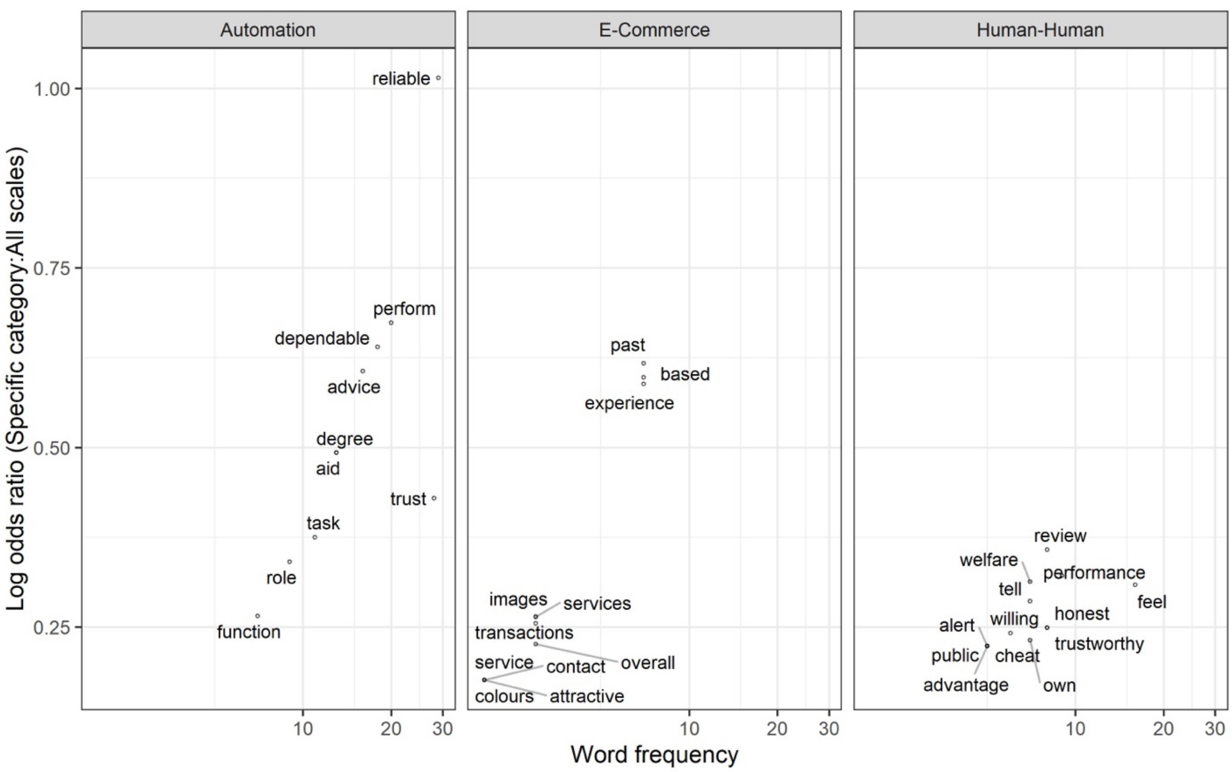

Figure 4. The log odds ratio of the 10 most common words in a specific category given all words in the questionnaires. The log odds ratio emphasizes words that are common in a specific category and relatively uncommon in others.

3.4 Questionnaire Composition

To help guide questionnaire selection, we assessed the questionnaire composition of domain-related words and layers of trust. The domain categories refer to trust in automation, e-commerce, or humans; the trust layers refer to dispositional, situational, or learned trust.

To calculate the proportion of domain-related words in each questionnaire (e.g., the questionnaire can have 20% automation-related terms, 30% e-commerce-related terms, and 50% human-human-related terms), we used the log odds ratio results shown in Figure 4. The questionnaires' domain composition results are illustrated in Figure 5, where the questionnaires are ordered by the proportion of automation, e-commerce, and human-human content. For instance, *Figure 5* shows that Goto's (1996) questionnaire is composed of 100% human-related words while the Delgado-Ballester, Munuera-Aleman & Yague-Guillen (2003) words involve a combination of each human, e-commerce, and automation-related words.

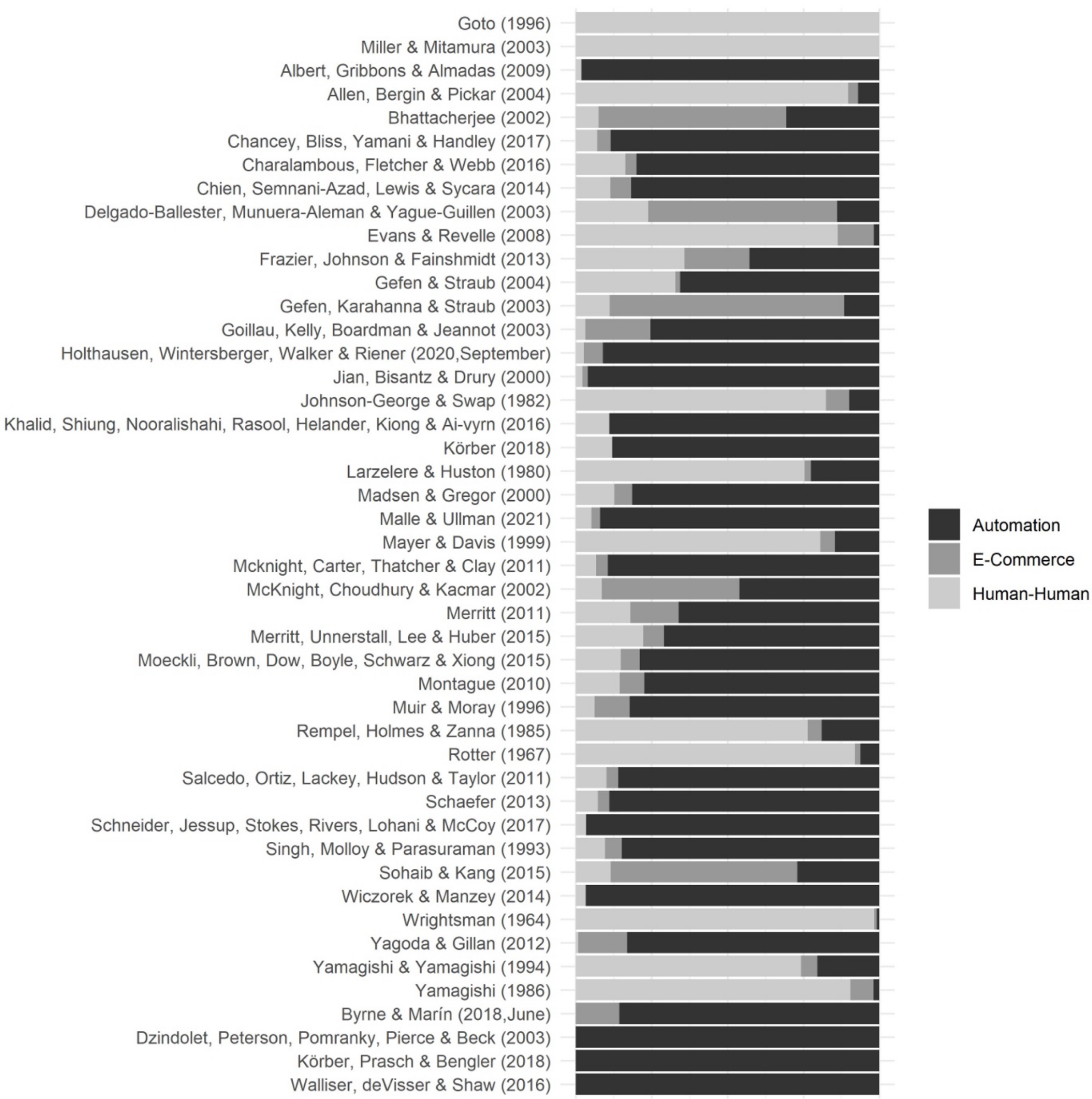

Figure 5. Trust questionnaires' domain composition (containing words most related to which domains of trust)

Similarly, we calculated the proportion of each of the trust layers for each item, which were coded by the authors, and the result is shown in Figure 6. Evans & Revelle (2008) is a questionnaire for measuring trust in humans and 100% of the questionnaire's items are learned; all items assess trust based on previous experiences. The Rotter (1967) questionnaire also measures trust in humans, but its items are 100% dispositional, meaning all items assess a person's innate tendency to trust. This is a starting point for selecting the questionnaire that best suits the research objective.

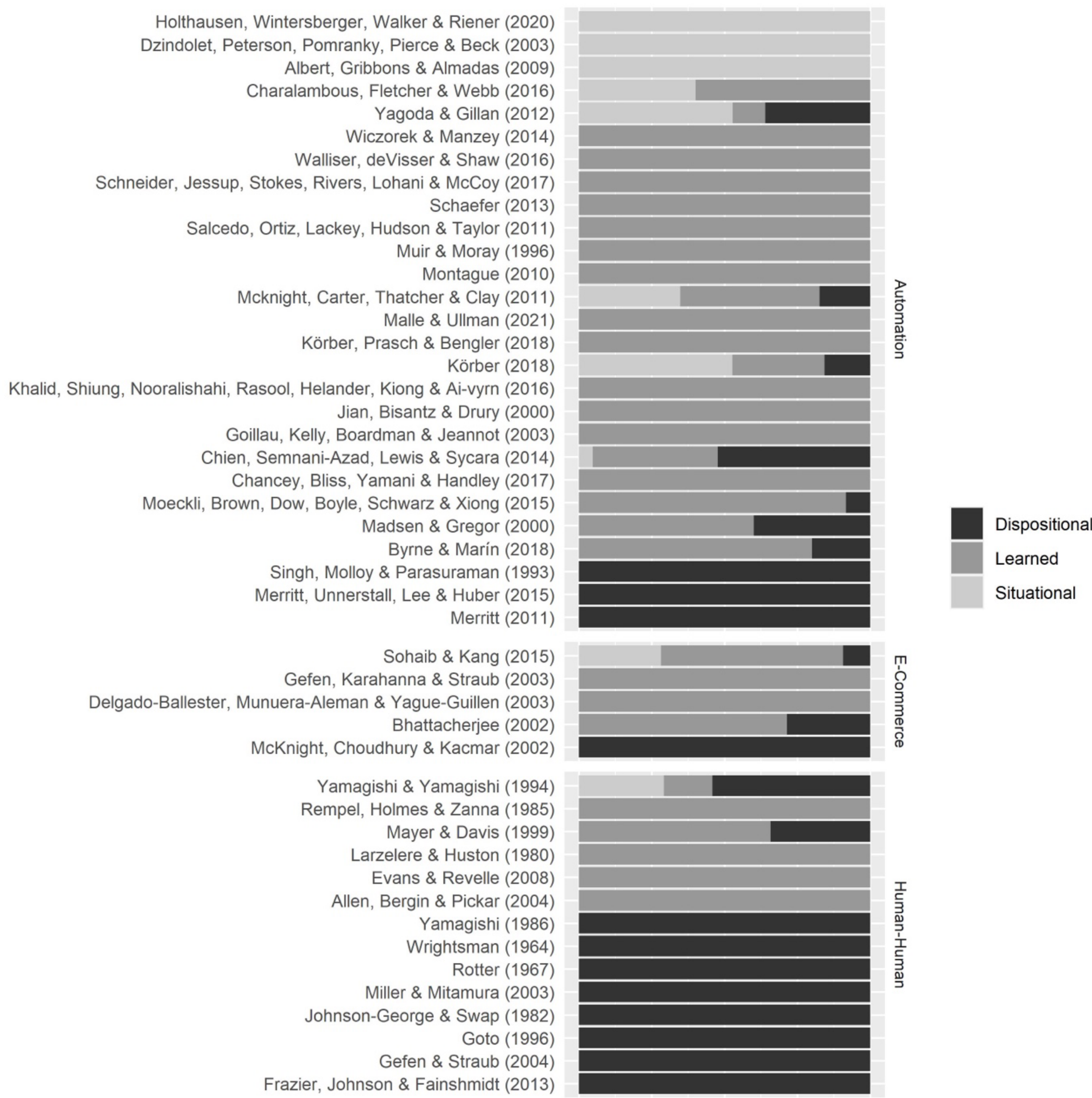

Figure 6. Trust questionnaires' layers of trust composition of each questionnaire item

3.5 Trust Lexicon

Figure 4 compared trust words in a specific domain category (i.e., automation, e-commerce, human-human) to trust words across all questionnaires. However, to identify trust-related words across domains, we compared them to the common words in the English language. We used a list of 5000 frequent words from the Corpus of Contemporary American English (COCA) (Davies & Gardner, 2010b). The corpus includes words from different genres; spoken, fiction, magazines, newspapers, and academic texts as shown in Figure *7*. Words above the dashed line ($y = 0$), indicate words that are more common in the trust questionnaires than in common English usage, while words below the dashed line are more prevalent in common English than in the trust questionnaires. Not surprisingly "trust" and "people" occur frequently in the questionnaires and occur much more frequently in the questionnaires than they do in English usage.

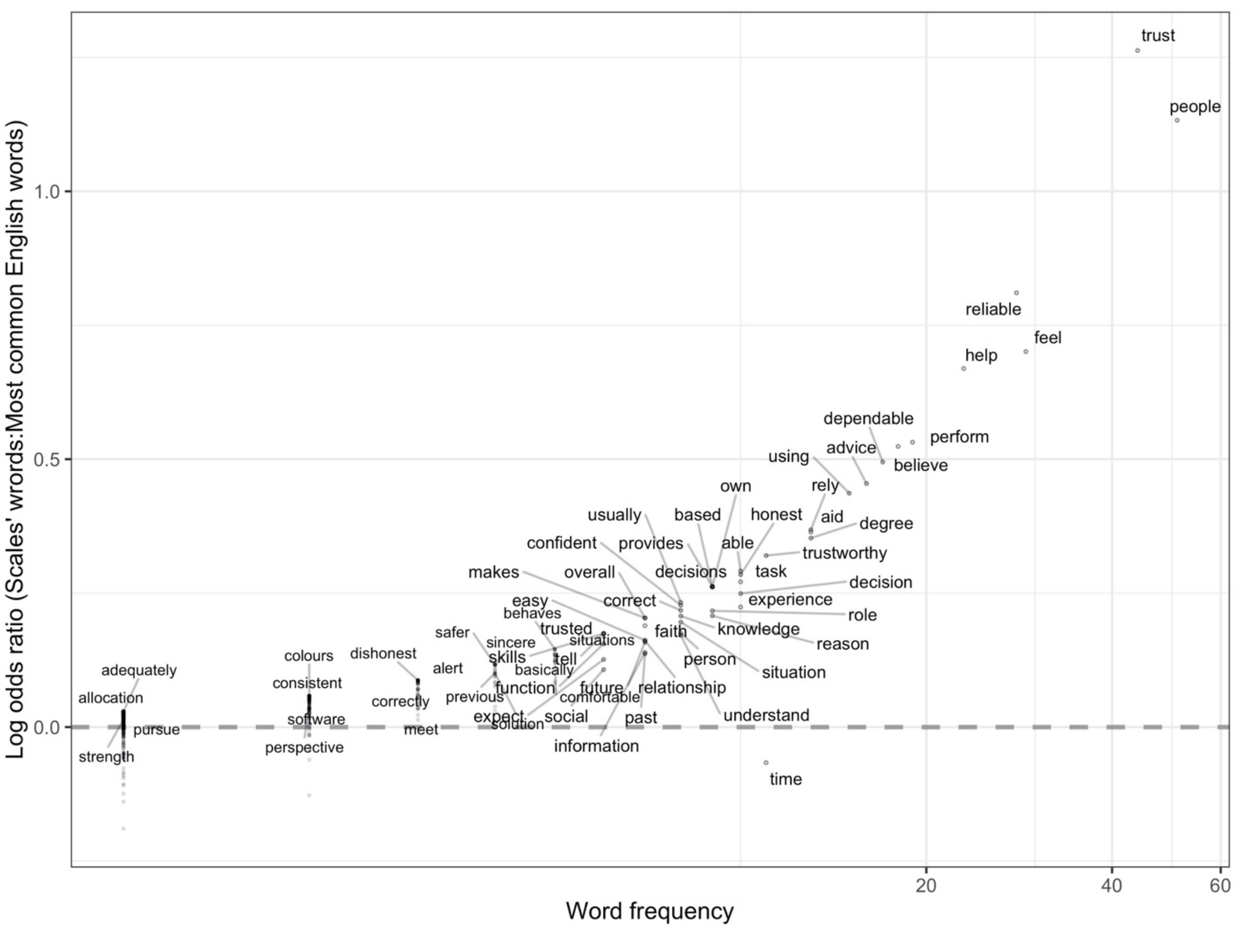

Figure 7. The log odds ratio of the words in the trust questionnaires to the words in the English language. The higher the word on the y-axis, the more unique the word is to trust questionnaires. Words to the right occur frequently in the trust questionnaires.

Based on the 20 most unique trust words revealed in Figure *7*, we created a trust lexicon shown in Figure 8. For each of the 20 words from Figure 7, we extracted the 20 closest neighboring words in the high-dimensional GloVe embeddings space using the cosine distance. Out of the 20 closest words, we selected the 10 most relevant ones. The 20 most unique trust words from the questionnaires are shown in light grey, while the other words in dark grey are their closest neighbors.

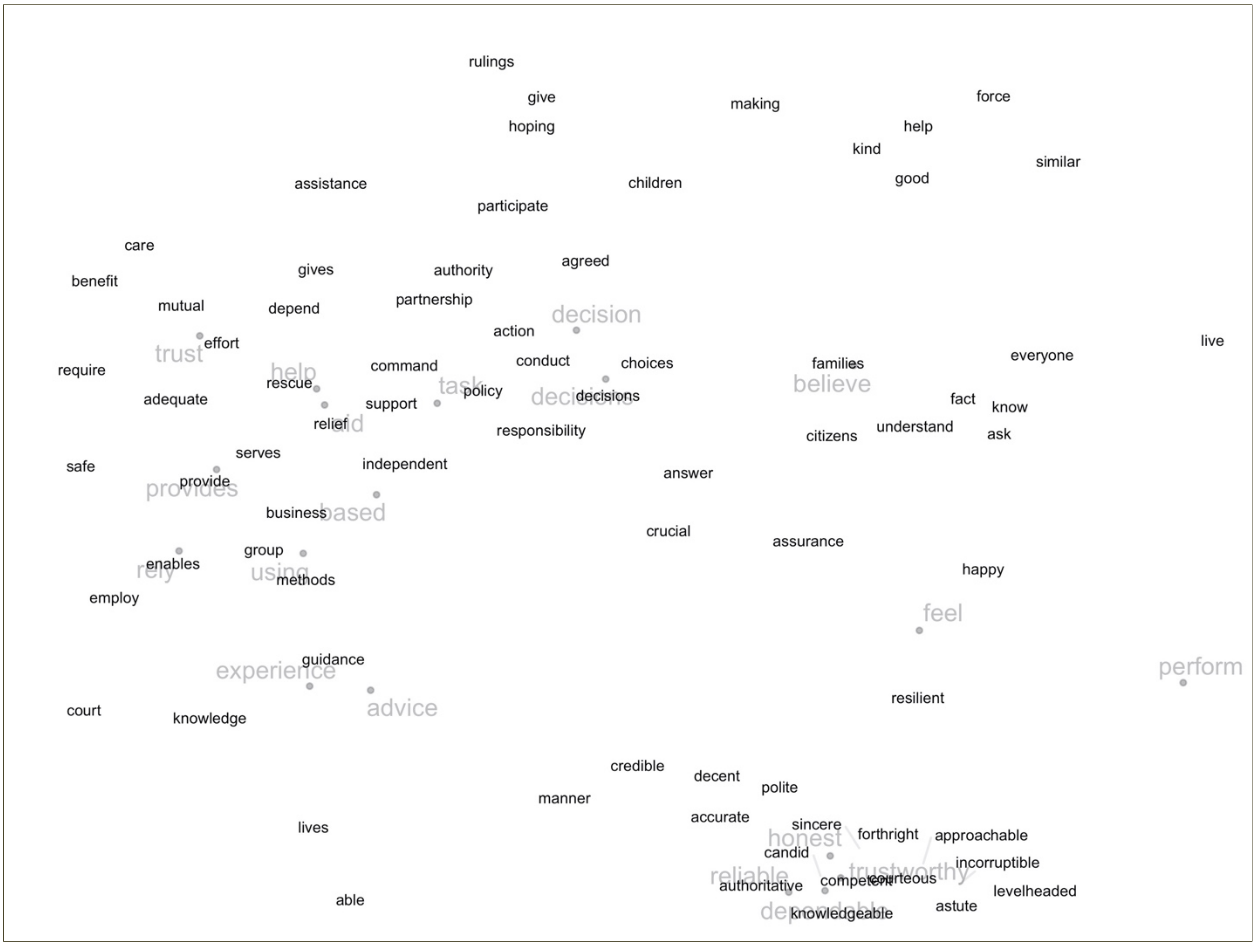

Figure 8. *A lexicon of trust-related words. The words in light grey are the most common in the trust questionnaires, while the rest of the words are their closest neighbors in the high-dimensional GloVe embeddings.*

4 Discussion

The results of the word-embedding text analysis identified the most common words used to assess trust, revealed semantic similarities and differences across the trust questionnaires, and provided a detailed comparison of the questionnaires' composition based on domain and trust layers. The results were implemented in an interactive web application that allows further exploration of the analysis. Particularly, the word-level and item-level results can support further questionnaire development and the questionnaire-level results can aid questionnaire selection given the research objective for measuring trust, and consideration of the trusting context.

4.1 Word-Level Analysis: Common Words and Trust Lexicon

The word-level analysis allowed us to calculate the frequently used words to measure trust. Common words included "dependable", "experience", and "advice". The neighboring words in the semantic space also had similar or complementary meanings such as “reliable”, “knowledge” and “helpful”, respectively. Words clusters shown in the figure can be mapped to trust dimensions identified by Malle & Ullman (2021). “Dependable” and “reliable” are closely situated next to each other, which can be mapped to the performance dimension of trust, whereas “secure” and “fair” show the moral dimension of trust. The themes revealed by the most frequent words were consistent with literature on trust dimensions (i.e., integrity and competence) (Malle & Ullman, 2021). When categorized into trust in Automation, Humans, and E-Commerce, a more specific theme emerged (e.g., reliability and dependability to describe automation. Honesty and sincerity to describe humans, and prior experiences and services to describe e-commerce.)

Understanding the words associated with trust questionnaires can be helpful in different ways. The identified trust-related words can be used as a single-word trust assessment tool, e.g., by asking subjects to rate how well these words describe the system of interest. The trust-related words can be descriptors used in card sorting tasks to measure trust measures, similar to the Microsoft Desirability Toolkit (Benedek & Miner, 2002). In addition, these words can be used as a basis for developing more robust or more precise instruments for measuring trust (Li, Alsaid, Noejovich, Cross, & Lee, 2020). By examining the most common words in each domain, we found similarities and differences in what questionnaires typically use to describe trust in automation compared to trust in humans. However, this should be interpreted carefully. Some forms of automation may be more human-like, such as anthropomorphic agents or virtual humans, hence using questionnaires involving more human-like qualities in these research contexts may be more appropriate (Lankton, Harrison Mcknight, & Tripp, 2015).

The words’ semantic space revealed by this analysis was used to create a trust lexicon. Word sentiment lexicons are typically created through tedious manual labeling of each word in the dictionary, which produces a sentiment rating for each word. But when word embeddings are combined with dimensionality reduction techniques, they reveal correlations between words and how they might relate to the overall construct of trust (Fast et al., 2017). The lexicon resulting from our analysis can build on and improve similar lexicons of trust-related words (Mohammad & Turney, 2013). Word lexicons can code people's usage to estimate people’s emotional state (Tausczik & Pennebaker, 2010) and attitudes on social media (Pang & Lee, 2008). Hence, creating trust-related lexicons can help with understanding peoples’ trust attitudes in different contexts and consequently help researchers construct the right tools to assess trust. For example, the trust lexicon could be used to score comments to identify instances of text or speech that relate to trust, similar to sentiment analysis (Mohammad & Turney, 2013; Pang & Lee, 2008).

4.2 Item-Level Analysis: Items’ Spread and Trust Characteristics

The item-level analysis showed that some questionnaires had items close to each other while items in other questionnaires were more dispersed. The distribution of the items was linked to the variation of the trust characteristics being evaluated by different items in each questionnaire (e.g., reliability, performance, cheating, prior experience, etc.); the greater the spread of a questionnaire’s items, the more dimensions are captured.

In general, the human-human questionnaires were broadly distributed. Because, in line with the questionnaire-level results, items in this domain tended to assess a wide range of human characteristics in different hypothetical scenarios. This implies that these questionnaires used words from varying contexts, which would explain their spread in the semantic space. On the other hand, some questionnaires’ items were contained in a very small area in the semantic space. One illustrative example is the most cited e-commerce questionnaire (Gefen et al., 2003); the questionnaire's items assess similar characteristics (and thus used closely related words) of online vendors such as "reliability", "honesty" and "trustworthiness".

In summary, the distribution of the questionnaire items in the semantic space can reflect the variety of trust dimensions being measured – the more dimensions the trust questionnaire captures, the more spread the items are. Therefore, one important consideration when selecting a trust questionnaire is the spread of its items. If the research question requires evaluating a specific quality that is associated with trust (e.g., reliability), then researchers could pick questionnaires with items that are closer together in the semantic space, whereas if the research questions require evaluating multiple qualities associated with trust (e.g., prior experience, performance, deception), a more spread questionnaire is likely more appropriate. Careful examination of the questionnaire and its constituent items is necessary.

4.3 Questionnaire-Level Analysis: Domain- and Layer- based Selection

At the questionnaire level, the results revealed three main clusters, one consisting of mainly human-human trust questionnaires, and two containing a mix of trust in automation and trust in e-commerce questionnaires. Questionnaires assessing a person's trust in other people were typically broad and contained diverse items assessing learned, dispositional, and situational trust through hypothetical scenarios, general views about the world, and the overall tendency to trust others. Questionnaires assessing trust in automation had two common themes with those assessing trust in e-commerce. Because after removing domain-specific words (e.g., "website" and "vendor" for e-commerce and "robot" and "automation" for automation), items of both domains were largely similar. One theme focused on assessments of reliability, accuracy, and trustworthiness of a system, while the other theme focused on the general tendency of people to trust or not trust new technologies. This explains the proximity in the semantic space, nonetheless, questionnaires developed for trust in automation might not be appropriate for to assess trust in e-commerce and vice versa. Depending on the context and the research question, one theme or a combination of themes might be more appropriate and researchers should carefully consider the aspects of trust being evaluated by each questionnaire (Kohn, de Visser, Wiese, Lee, & Shaw, 2021).

4.4 Practical Implications: Questionnaire Selection Guidelines

The questionnaire composition analysis provided further clarity on how the myriad trust questionnaires compare to one another and can thus serve as initial guidance for selecting a questionnaire. Here, we outline general guidelines and considerations for the trust questionnaire selection process: identifying the domain and layer, and considering items' dispersion, and evaluating the tradeoff between number of items and sampling frequency.

After carefully defining the research questions and the underlying hypotheses, the researcher needs to identify the domain in which trust is being measured. This is important because trust questionnaires are typically developed to measure trust in a specific context and the way trust is characterized varies from one domain to another (Lee & See, 2004; Lewis & Weigert, 2012). This was evident in the word-level analysis in Figure 4: the words used to describe trust differed across domains. For example, Lee and Moray's (1992) conceptualization of purpose (e.g., role), process (e.g., dependable), and performance (e.g., reliable) dimensions of trust in automation was apparent in some of the words, as well as the concept that trust may also have more moral dimensions (Mayer et al., 1995; Sheridan, 2019). Furthermore, in Figure 5, we provided a questionnaire composition map to show what percentage of each questionnaire included words most unique to trust in automation, human-human, or e-commerce. Nonetheless, this was based on the objective quantitative analysis, researchers should carefully assess whether or not a questionnaire is appropriate for measuring trust in a certain domain.

Moreover, identifying the *layer* of trust is important; whether the researcher is trying to assess people's general propensity to trust (i.e., dispositional), trust in a specific situation (i.e., situational), or trust based on previous experiences (i.e., learned). In Figure 6, we provided a map for understanding the composition of the questionnaires. Based on the research questions, the selected questionnaire items can be dispositional, learned, situational, or a combination (Merritt & Ilgen, 2008). It is important in this step to understand the nature of each trust layer. Measuring dispositional trust would be most appropriate for studies of individual differences, particularly when measuring trust across different cultures, as people from different cultures may have different perceptions of trust. Moreover, learned trust would be for studies of how interactions with an agent affect trust, and situational trust would be for measuring trust in a specific event. For example, if researchers are interested in evaluating the users' trust in automation in specific of interactions, a questionnaire that mainly consists of situational trust items would be suitable (e.g., Holthausen, Wintersberger, Walker & Riener, 2020) whereas if they wish to assess the persons' propensity to trust, a questionnaire like that mainly consists of dispositional trust item (e.g., Meritt (2011)) would be better suited. Nonetheless, the trust layer is highly dependent on context. The same question can be used to assess different layers of trust depending on when it is administered (i.e., before, during, or after a situation) (Merritt & Ilgen, 2008).

Furthermore, the results revealed another element of the questionnaires' semantic characteristics and selection criterion: questionnaire items spread. The *spread* of the questionnaire items is an important criterion of selection: whether the research question and nature of the study focus on one or a few of the dimensions of trust (e.g., purpose, process, or performance information for forming a person's trust in automation (Lee & See, 2004), or rational and relational dimensions of trust in others (Lewis & Weigert, 2012)). This can be qualitatively determined by visually assessing the specific questionnaire items' distribution across the semantic space, or quantitatively by the spread values in Table 2 that were calculated as the mean of Euclidian distances from a questionnaire's centroid in the semantic space. When selecting a questionnaire, if researchers are interested in trust as a moderator or control variable (Fuchs & Diamantopoulos, 2009) or only focusing on a single aspect of trust (e.g., the performance of a particular automated system), then picking a narrower spread of trust scale can be appropriate (e.g., Chancey, Bliss, Yamani & Handley (2017) for trust in automation which focuses on the ability and dependability dimensions of trust , Goto (1996) for trust in humans which measures trust relative to social distance, or McKnight, Choudhury & Kacmar (2002) for trust in e-commerce which measures tendency in particular.) If researchers are interested in trust as an essential variable in the study, then a broader spread of trust scale should be considered (e.g., Schaefer (2013) for trust in automation, Mayer & Davis (1999) for trust in humans, or Sohaib & Kang (2015) or e-commerce).

In addition, evaluating the trade-off between the *number of questionnaire items* and *sampling frequency* is critical (Kohn, de Visser, Wiese, Lee, & Shaw, 2021). If trust needs to be measured multiple times for its dynamic characteristic, using a few or single itemed questionnaire might provide a quick trust measurement and minimal interruptions to the continuity of a study participant's experience (e.g., Körber (2018)). However, one item might be limited and not measure the different dimensions of trust (Lee & See, 2004). If the research objective requires a more detailed assessment of trust, then multi-item questionnaires are recommended (e.g., Yagoda & Gillan (2012)). This is particularly important in situations where different layers of trust need to be measured at different times of a study (i.e., dispositional trust before the study, situational trust during the study.). In combination with the questionnaire composition analysis, the researcher can make an informed decision regarding the questionnaire selection with the right number of questions that meets the research needs.

Finally, the supplemented web app implementation provides an interactive interface to compare, contrast and select the questionnaire most appropriate based on the considerations provided above. For a more detailed explanation and description of how to use the app for questionnaire exploration and selection see Appendix B.

4.5 Limitations and Further Research

This study has several limitations. First, some of the identified words as part of the trust lexicon (e.g., "feel" or "believe") may be more an artifact of the measurement method and our ability to elicit self-report from lay-people (i.e., not trust scholars) through a scaled question (e.g., "how much do you feel…"), rather than an indicator of what trust is, which has been defined as an attitude, and not an intention, feeling, or belief (Lee & See, 2004). These and other words may reflect how the question was framed rather than the question content.

Second, in the questionnaire composition analysis, we show what each questionnaire measures, in terms of domain and layers of trust, while remaining agnostic as to whether or not these questionnaires measure them well. Researchers should self-assess and investigate further the validity of each measure for their research task at hand, as is the standard practice of scientific rigor. Furthermore, the questionnaire composition analysis does not precisely reflect the effect of the number of items. That is, a single-item questionnaire would be 100% dispositional, learned, or situational. But that does not necessarily mean that it is the best questionnaire to measure that specific trust layer. In addition, the trust layer categorization was based on the specific questionnaire's purpose, however, the same question can be used to assess different layers of trust depending on the context and the time it was administered (i.e., before, during, or after an interaction) (Merritt & Ilgen, 2008).

Third, our categorization of the trust questionnaires' domains is rather generic. For example, some automation trust questionnaires are targeted at trust in automated vehicles specifically whereas others are targeted at trust in automation in general. We tackle this in the analysis by removing system-specific terms (e.g., robot, vehicle, website), however, questionnaire specificity remains an important consideration in questionnaire selection. A questionnaire that assesses trust in an assistive robot might have more human-like questions that do not necessarily translate to trust in an automated vehicle.

Fourth, the similarity found between the trust in automation and trust in e-commerce questionnaires may have been due to having similar theoretical origins; both categories assess trust in some type of technology or trust in an entity mediated by technology (Ghazizadeh, Lee, & Boyle, 2012) – to understand trust-related decisions such as reliance or purchasing. We are not claiming that this is a novel finding, and indeed assessing the history of these questionnaires would lead to a similar insight. However, our approach reveals this relationship through a quantitative, systematic analysis that shows researchers across multiple domains have similarly operationalized the construct of trust.

Fifth, our approach to characterizing the similarities and differences between questionnaires was data-driven. Data-driven approaches have been proven useful in expanding knowledge and extracting scientific relationships years in advance of their discovery (Tshitoyan et al., 2019). Yet, trust is a complex, multifaceted construct and future work should incorporate a theory-driven approach to safeguard the theoretical underpinnings of trust, expand trust theory, and build on existing measures (Long et al., 2020; McCroskey & Young, 1979).

Sixth, we are potentially missing some questionnaires – a more comprehensive review might have revealed more relevant questionnaires, such as studies that focus on information credibility that may be related to trust (Fogg & Tseng, 1999; McCroskey & Teven, 1999; Renn & Levine, 1991). However, this paper demonstrates a scalable method that could be easily applied to an expanded corpus if new questionnaires are developed or to explore a broader conception of trust. In addition, the study included questionnaires in the corpus that were developed ad hoc, relying primarily on face validity, or were not empirically validated. This may lead to potential issues for item selection.

Finally, one important limitation of this study is that it only included questionnaires developed in English. Excluding trust questionnaires that might have been developed in another language has important implications for advancing trust theory and methods of trust measurement across languages and cultures, but also for generalizability of the trust lexicon.

5 Conclusion

This study demonstrates the potential of text analysis in understanding and assessing trust in different contexts, which provides a systematic method to quantify similarities and differences for further survey development and questionnaire selection.

The analyses conducted were at the word, item, and questionnaire levels. Each highlighted important considerations of questionnaire development and selection. The word-level analysis showed the most common words and themes that emerged from the trust questionnaires literature and produces a trust lexicon. This has implications for questionnaire development and understanding of trust in conversational speech and public attitudes on social media (Li et al., 2020; Pang & Lee, 2008). Furthermore, the item and questionnaire analyses provided higher-level insights into questionnaire items composition, and questionnaire items spread across the semantic space, both of which are important considerations for questionnaire selection.

While this study focused on text-based trust questionnaires, this approach can be extended to more specific domains; such as estimating drivers' trust in self-driving vehicles through speech, similar to previous work on emotion classification (Aman & Szpakowicz, 2007) and the analysis of open-ended survey responses (Lee & Kolodge, 2018).

Overall, word-embedding text analysis is a useful way to understand the sentiments and emotions associated with words. The resulting semantic space of trust words provides a way to compare and select trust questionnaires. In addition, the resulting lexicon of trust-related words can be used in natural language processing to understand trust attitudes through conversations between people, and between people and technologies in different domains.

6 References

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Upper Saddle River, NJ: Prentice Hall.

Albert, W., Gribbons, W., & Almadas, J. (2009). Pre-conscious assessment of trust: A case study of financial and health care web sites. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *53*(6), 449–453. doi:10.1177/154193120905300603

Allen, K., Bergin, R., & Pickar, K. (2004). Exploring trust, group satisfaction and performance in geographically dispersed and co-located university technology commercialization teams. *Education that Works: The NCIIA 8th Annual Meeting, March 18-20, 2004*.

Alsaid, A. (2020). Estimating Driver State in Increasingly Automated Vehicles, The University of Wisconsin-Madison.

Alsaid, A., & Lee, J. D. (2022). The DataScope: A mixed-initiative architecture for data labeling. *Human Factors and Ergonomics Society Annual Meeting*.

Alsaid, A., Lee, J. D., Roberts, D. M., Barrigan, D., & Baldwin, C. L. (2018). Looking at mind wandering during driving through the windows of PCA and t-SNE. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*(1), 1863–1867. SAGE PublicationsSage CA: Los Angeles, CA. doi:10.1177/1541931218621424

Altszyler, E., Sigman, M., Ribeiro, S., & Slezak, D. F. (2016). Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database. doi:10.1016/j.concog.2017.09.004

Aman, S., & Szpakowicz, S. (2007). Identifying expressions of emotion in text. *Text, Speech and Dialogue* (pp. 196–205). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-540-74628-7_27

Benedek, J., & Miner, T. (2002). Measuring desirability: New methods for evaluating desirability in a usability lab setting. *Proceedings of Usability Professionals Association* (pp. 8–12).

Bhattacherjee, A. (2002). Individual trust in online firms: Scale development and initial test. *Journal of Management Information Systems*, *19*(1), 211–241. doi:10.1080/07421222.2002.11045715

Byrne, K., & Marín, C. (2018). Human trust in robots when performing a service. *2018 IEEE 27th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105. doi:10.1037/h0046016

Cassell, J., & Bickmore, T. (2000). External manifestations of trustworthiness in the interface. *Communications of the ACM*, *43*(12), 50–56. doi:10.1145/355112.355123

Chancey, E. T., Bliss, J. P., Yamani, Y., & Handley, H. A. H. (2017). Trust and the compliance-reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence. *Human Factors*, *59*(3), 333–345. doi:10.1177/0018720816682648

Charalambous, G., Fletcher, S., & Webb, P. (2016). The development of a scale to evaluate trust in industrial human-robot collaboration. *International Journal of Social Robotics*, *8*(2), 193–209. doi:10.1007/s12369-015-0333-8

Chien, S.-Y., Semnani-Azad, Z., Lewis, M., & Sycara, K. (2014). Towards the ddevelopment of an inter-cultural scale to measure trust in automation. *International Conference on Cross-Cultural Design* (pp. 35–46). doi:10.1007/978-3-319-07308-8_4

Chiou, E. K., & Lee, J. D. (2021). Trusting automation: Designing for responsivity and resilience. *Human Factors*.

Chita-Tegmark, M., Law, T., Rabb, N., & Scheutz, M. (2021). Can you trust your trust measure? *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (Vol. 4, pp. 92–100). New York, NY, USA: ACM. doi:10.1145/3434073.3444677

Davies, M., & Gardner, D. (2010a). Word Frequency. *Word Frequency List of American English*. Retrieved a from https://www.wordfrequency.info/free.asp

Davies, M., & Gardner, D. (2010b). Word Frequency. *Word Frequency List of American English*. Retrieved b from

Delgado-Ballester, E., Munuera-Aleman, J. L., & Yague-Guillen, M. J. (2003). Development and validation of a brand trust scale. *International Journal of Market Research*, *45*(1), 35–54. doi:doi.org/10.1177/147078530304500103

Domeyer, J., Venkatraman, V., Price, M., & Lee, J. D. (2018). Characterizing Driver Trust in Vehicle Control Algorithm Parameters. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*(1), 1821–1825. doi:10.1177/1541931218621413

Dzindolet, M., Peterson, S., Pomranky, R., Pierce, L., & Beck, H. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718. doi:10.1016/S1071-5819(03)00038-7

Evans, A. M., & Revelle, W. (2008). Survey and behavioral measurements of interpersonal trust. *Journal of Research in Personality*, *42*(6), 1585–1593. Elsevier Inc. doi:10.1016/j.jrp.2008.07.011

Fast, E., Chen, B., & Bernstein, M. S. (2017). Lexicons on demand: Neural word embeddings for large-scale text analysis. *IJCAI International Joint Conference on Artificial Intelligence*. doi:10.24963/ijcai.2017/677

Fogg, B. J., & Tseng, H. (1999). The elements of computer credibility. *Proceedings of the SIGCHI conference on Human factors in computing systems the CHI is the limit - CHI ’99* (pp. 80–87). New York, New York, USA: ACM Press. doi:10.1145/302979.303001

Frazier, M. L., Johnson, P. D., & Fainshmidt, S. (2013). Development and validation of a propensity to trust scale. *Journal of Trust Research*, *3*(2), 76–97. doi:10.1080/21515581.2013.820026

Fuchs, C., & Diamantopoulos, A. (2009). Using single-item measures for construct measurement in management research: Conceptual issues and application guidelines. *Die Betriebswirtschaft*.

Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and tam in online shopping: An integrated model. *MIS*, *27*(1), 51–90. doi:10.2307/30036519

Gefen, D., & Straub, D. W. (2004). Consumer trust in B2C e-Commerce and the importance of social presence: Experiments in e-Products and e-Services. *Omega*, *32*(6), 407–424. doi:10.1016/j.omega.2004.01.006

Ghazizadeh, M., Lee, J. D., & Boyle, L. N. (2012). Extending the technology acceptance model to assess automation. *Cognition, Technology & Work*, *14*(1), 39–49. doi:10.1007/s10111-011-0194-3

Goillau, P., Kelly, C., Boardman, M., & Jeannot, E. (2003). *Guidelines for trust in future atm systems-measures*.

Goto, S. G. (1996). To trust or not to trust: Situational and dispositional determinants. *Social Behavior and Personality*, *24*(2), 119–131. doi:10.2224/sbp.1996.24.2.119

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal*, *26*, 91–109. doi:10.1111/j.1471-1842.2009.00848.x

Hoff, K. A., & Bashir, M. (2014). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*. doi:10.1177/0018720814547570

Holthausen, B. E., Wintersberger, P., Walker, B. N., & Riener, A. (2020). Situational Trust Scale for Automated Driving (STS-AD): Development and Initial Validation. *International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 40–47). doi:10.1145/3409120.3410637

Hu, W., Zhang, J., & Zheng, N. (2016). Different contexts lead to different word embeddings. *COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers*.

Hubert, M., Rousseeuw, P. J., & Vanden Branden, K. (2005). ROBPCA: A new approach to robust principal component analysis. *Technometrics*, *47*(1), 64–79. doi:10.1198/004017004000000563

Jeong, H., Park, J., Park, J., Pham, T., & Lee, B. C. (2018). Analysis of trust in automation survey instruments using semantic network analysis. *International conference on applied human factors and ergonomics* (pp. 9–18).

Jian, J.-Y., Bisantz, A. M., Drury, C. G., & Llinas, J. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71. doi:10.1207/S15327566IJCE0401_04

Johnson-George, C., & Swap, W. C. (1982). Measurement of specific interpersonal trust: Construction and validation of a scale to assess trust in a specific other. *Journal of Personality and Social Psychology*, *43*(6), 1306. doi:10.1037/0022-3514.43.6.1306

Khalid, H. M., Shiung, L. W., Nooralishahi, P., Rasool, Z., Helander, M. G., Kiong, L. C., & Ai-Vyrn, C. (2016). Exploring psycho-physiological correlates to trust: Implications for human-robot-human interaction. *Proceedings of the Human Factors and Ergonomics Society* (pp. 696–700). Human Factors an Ergonomics Society Inc. doi:10.1177/1541931213601160

Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y.-C., & Shaw, T. H. (2021). Measurement of trust in automation: A narrative review and reference guide. *Frontiers in Psychology*, *12*. doi:10.3389/fpsyg.2021.604977

Konopka, T. (2019). Uniform Manifold Approximation and Projection.

Körber, M. (2018). Theoretical considerations and development of a questionnaire to measure trust in automation. *Congress of the International Ergonomics Association* (Vol. 823, pp. 13–30). doi:10.1007/978-3-319-96074-6_2

Körber, M., Prasch, L., & Bengler, K. (2018). Why do I have to drive now? Post hoc explanations of takeover requests. *Human Factors*, *60*(3), 305–323. SAGE Publications Inc. doi:10.1177/0018720817747730

Landauer, T. K., & Dumais, S. (2008). Latent semantic analysis. *Scholarpedia*, *3*(11), 4356.

Lankton, N. K., Harrison Mcknight, D., & Tripp, J. (2015). Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems*, *16*(10), 880–918. doi:10.17705/1jais.00411

Larzelere, R., & Huston, T. (1980). The dyadic trust scale: Toward understanding interpersonal trust in close relationships. *Journal of Marriage and the Family*, *42*(3), 595–604.

Lee, J. D., & Kolodge, K. (2018). Understanding Attitudes Towards Self-Driving Vehicles: Quantitative Analysis of Qualitative Data. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*(1), 1399–1403. doi:10.1177/1541931218621319

Lee, J. D., Liu, S.-Y., Domeyer, J., & DinparastDjadid, A. (2019). Assessing drivers' trust of automated vehicle driving styles with a two-part mixed model of intervention tendency and magnitude. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 001872081988036. doi:10.1177/0018720819880363

Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–1270. doi:10.1080/00140139208967392

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80. doi:10.1518/hfes.46.1.50.30392

Lewis, J. D., & Weigert, A. J. (2012). The social dynamics of trust: Theoretical and empirical research, 1985-2012. *Social Forces*, *91*(1), 25–31. doi:10.1093/sf/sos116

Li, M., Alsaid, A., Noejovich, S. I., Cross, E. V., & Lee, J. D. (2020). Towards a Conversational Measure of Trust. Retrieved from http://arxiv.org/abs/2010.04885

Long, S. K., Sato, T., Millner, N., Loranger, R., Mirabelli, J., Xu, V., & Yamani, Y. (2020). Empirically and theoretically driven scales on automation trust: A multi-level confirmatory factor analysis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *64*(1), 1829–1832. doi:10.1177/1071181320641440

Maaten, L. Van Der, & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research 1*, *620*(1), 267–84. doi:10.1007/s10479-011-0841-3

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. *Proceedings of Eleventh Australasian Conference on Information Systems* (pp. 6–8).

Malle, B. F., & Ullman, D. (2021). A multidimensional conception and measure of human-robot trust. *Trust in Human-Robot Interaction* (pp. 3–25). Elsevier. doi:10.1016/B978-0-12-819472-0.00001-0

Mayer, R. C., & Davis, J. H. (1999). The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of Applied Psychology*. doi:10.1037/0021-9010.84.1.123

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *The Academy of Management Review*, *20*(3), 709–734. doi:10.2307/258792

McCroskey, J. C., & Teven, J. J. (1999). Goodwill: A reexamination of the construct and its measurement.

Communication Monographs, *66*(1), 90–103. doi:10.1080/03637759909376464

McCroskey, J. C., & Young, T. J. (1979). The use and abuse of factor analysis in communication research. *Human Communication Research*, *5*(4), 375–382. doi:10.1111/j.1468-2958.1979.tb00651.x

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*. doi:10.21105/joss.00861

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems*, *2*(2). doi:10.1145/1985347.1985353

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, *13*(3), 334–359. doi:10.1287/isre.13.3.334.81

Merritt, S. M. (2011). Affective processes in human-automation interactions. *Human Factors*, *53*(4), 356–370. doi:10.1177/0018720811411912

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors*, *50*(2), 194–210. doi:10.1518/001872008X288574

Merritt, S. M., Unnerstall, J. L., Lee, D., & Huber, K. (2015). Measuring individual differences in the perfect automation schema. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(5), 740–753. doi:10.1177/0018720815581247

Meyer, J., & Lee, J. D. (2013). Trust, reliance, and compliance. *The Oxford Handbook of Cognitive Engineering*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations ofwords and phrases and their compositionality. *Advances in Neural Information Processing Systems*.

Miller, A. S., & Mitamura, T. (2003). Are surveys on trust trustworthy? *Social Psychology Quarterly*, *66*(1), 62.

Moeckli, J., Brown, T., Dow, B., Boyle, L. N., Schwarz, C., & Xiong, H. (2015). *Evaluation of adaptive cruise control interface requirements on the national advanced driving simulator*.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*. doi:10.1111/j.1467-8640.2012.00460.x

Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, *16*(4), 372–403. doi:10.1093/pan/mpn018

Montague, E. (2010). Validation of a trust in medical technology instrument. *Applied ergonomics*, *41*(6), 812–21. Elsevier Ltd. doi:10.1016/j.apergo.2010.01.009

Muir, B., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*(3), 429–460.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*, 1–135. doi:10.1561/1500000001

Patel, F. N. (2016). Large high dimensional data handling using data reduction. *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)* (pp. 1531–1536). IEEE. doi:10.1109/ICEEOT.2016.7754940

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. doi:10.3115/v1/d14-1162

Price, M., Lee, J. D., Dinparastdjadid, A., Toyoda, H., & Domeyer, J. (2017). Effect of vehicle control algorithms on eye behavior in highly automated vehicles. *FAST-zero*.

R Development Core Team. (2016). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*. doi:10.1007/978-3-540-74686-7

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). *Trust in close relationships*. *Review of Personality and Social Psychology*.

Renn, O., & Levine, D. (1991). Credibility and trust in risk communication. *Communicating Risks to the Public* (pp. 175–217). Dordrecht: Springer Netherlands. doi:10.1007/978-94-009-1952-5_10

Rotter, J. B. J. (1967). A new scale for the measurement of interpersonal trust. *Journal of personality*, *35*(4), 651–65. doi:10.1111/j.1467-6494.1967.tb01454.x

Salcedo, J. N., Ortiz, E. C., Lackey, S. J., Hudson, I., & Taylor, A. H. (2011). Effects of autonomous vs. remotely-operated unmanned weapon systems on human-robot teamwork and trust. *Proceedings of the Human Factors and Ergonomics Society* (pp. 635–639). doi:10.1177/1071181311551130

Schaefer, K. E. (2013). *The perception and measurement of human-robot trust*.

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience. *The Behavioral and Brain Sciences*, *36*, 393–414. doi:10.1017/S0140525X12000660

Schneider, T. R., Jessup, S. A., C., S., Rivers, S., Lohani, M., & McCoy, M. (2017). The influence of trust propensity on behavioral trust. *Poster session presented at the meeting of Association for Psychological Society,*.

Sheridan, T. B. (2019). Individual differences in attributes of trust in automation: Measurement and application to system design. *Frontiers in Psychology*, *10*(1117). doi:10.3389/fpsyg.2019.01117

Silge, J., & Robinson, D. (2016). Text Mining with R. *The Journal of Open Source Software*. doi:10.21105/joss.00037

Singh, I. L., Molloy, R., & Parasuraman, R. (1993). Automation-induced "complacency": Development of the complacency-potential rating scale. *The International Journal of Aviation Psychology*, *3*(2), 111–122. doi:10.1207/s15327108ijap0302_2

Sohaib, O., & Kang, K. (2015). Individual level culture influence on online consumer iTrust aspects towards purchase intention across cultures: A SOR model. *International Journal of Electronic Business*, *12*(2), 142–161. doi:10.1504/IJEB.2015.069104

Takayama, L. (2009). Making sense of agentic objects and teleoperation: In-the-moment and reflective perspectives. *Human-Robot Interaction (HRI), 2009 4th ACM/IEEE International Conference on*, 239–240. doi:10.1145/1514095.1514155

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*. doi:10.1177/0261927X09351676

Walliser, J. C., De Visser, E. J., & Shaw, T. H. (2016). Application of a system-wide trust strategy when supervising multiple autonomous agents. *Proceedings of the Human Factors and Ergonomics Society* (pp. 133–137). doi:10.1177/1541931213601031

Wickham, H. (2016). *tidyverse: Easily Install and Load "Tidyverse" Packages. R package version 1.0.0.*

Wickham, H., & Winston, C. (2019). Create Elegant Data Visualisations Using the Grammar of Graphics. *Package "ggplot2."* doi:10.1093/bioinformatics/btr406

Wiczorek, R., & Manzey, D. (2014). Supporting attention allocation in multitask environments: Effects of likelihood alarm systems on trust, behavior, and performance. *Human Factors*, *56*(7), 1209–1221. SAGE Publications Inc. doi:10.1177/0018720814528534

Wrightsman, L. S. (1964). Measurement of Philosophies of Human Nature. *Psychological Reports*, *14*(3), 743–751. doi:10.2466/pr0.1964.14.3.743

Yagoda, R. E., & Gillan, D. J. (2012). You want me to trust a ROBOT? The development of a human-robot interaction trust scale. *International Journal of Social Robotics*, *4*(3), 235–248. Springer. doi:10.1007/s12369-012-0144-0

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, *51*(1), 110–116. doi:10.1037/0022-3514.51.1.110

Yamagishi, T., & Yamagishi, M. (1994). Trust and commitment in the United States and Japan. *Motivation and Emotion*, *18*(2), 129–166. doi:10.1007/BF02249397

Yang, X. J., Schemanske, C., & Searle, C. (2021). Toward quantifying trust dynamics: How people adjust their trust after moment-to-moment interaction with automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 001872082110347. doi:10.1177/00187208211034716