Survey Research Methods (2024)
Vol. 18, No. 2, pp. 153-165
https://doi.org/10.18148/srm/2024.v18i2.8240
European Survey Research Association

©2024 Author(s)
CC BY 4.0

Evaluating Item Content and Scale Characteristics Using a Pretrained Neural Network Model

Jeffrey Stanton[1] · Angela Ramnarine-Rieks[1] · Yisi Sang[1]
[1]Syracuse University

Multi-item scales are widely used in social research. The psychometric characteristics of a scale and the successful use of a scale in research depend in part on item wording. This article demonstrates a method for using natural language processing (NLP) tools to assist with the item development process, by showing that numeric embedding representations of items are useful in predicting the characteristics of a scale. NLP comprises a set of algorithmic techniques for analysing words, phrases, and larger units of written language. We used NLP tools to create and analyse semantic summaries of the item texts for n=386 previously published multi-item scales. Results showed that semantic representations of items connect to scale characteristics such as Cronbach's alpha internal consistency.

Keywords: Cronbach's alpha; Answer behavior; Emotion prediction; Natural Language Processing; Open-ended questions; Neural network; Rating scale

1 Introduction

In 1932, Rensis Likert published, "A technique for the measurement of attitudes," which simplified earlier scaling procedures pioneered by Louis Thurstone (Edmondson, 2005). The so-called Likert-scale enabled researchers from many fields to construct items and scales using straightforward techniques. In the ensuing nine decades, researchers from many social science areas have written, field tested, and published validation data for many such attitude scales. A variety of researchers have examined approaches and options for constructing these items, and this work has been ably synthesised into helpful advice by Jebb, Ng, and Tay, 2021; Calderón, Morales, Liu, and Hays, 2006; Clark and Watson, 1995, 2019, and others. Advice for writing items for multi-item scales generally focuses on using language familiar to the intended respondents, promoting readability, avoiding double-barrelled constructions, and balancing the specificity and generality of item content. Once written, evaluating the suitability of items for measuring the attitude of interest typically occurs through collection of pilot data from respondents. Psychometric analyses then indicate which items seem to function well and which function poorly, enabling refinement of the scale.

Corresponding author: Jeffrey Stanton, Syracuse University, Syracuse, NY, USA (Email: jmstanto@syr.edu)

Over recent years, new computational techniques have emerged for systematic analysis of the linguistic content of natural language text (Kobayashi, Mol, Berkers, Kismihók, & Den Hartog, 2018). The possibility thus arises that such techniques could enhance and complement the typical activities of psychometric analysis by adding tools for systematically investigating the language used in scale items. The present research begins the exploration of this idea. Little is known about whether or how the output of computational linguistic models connects with the conventional methods of scale development and analysis.

The main contribution of this research arises from showing that patterns of semantic relationships among items can connect to quantifiable characteristics of self-report scales. While computational linguistic analysis cannot substitute for psychometric analysis, it may provide opportunities for researchers to learn helpful information about items before, during, and after scale development.

2 Background

This literature review examines three main topics: research on item creation and analysis; techniques to represent the semantic content of natural language texts as high dimensional numeric vectors; and an overview of the predictive neural network model used in this article. The first section provides context for understanding how recommended item development processes lead to specific collections of short

natural language texts, as well as how item content connects to the characteristics and usage of the resulting scales. The second section outlines developments in creating vector representations of text culminating with an introduction to how to produce numeric representations of semantics. The last section describes essential aspects of the convolutional neural networks used as the primary analysis approach in the present study. Together these areas of literature review lead up to an analysis of a large collection of items and scales that attempts to connect the semantic content of the item texts to three quantities of interest: Cronbach's alpha reliability of a scale as reported in a validation article and two measures pertaining to the validation article itself.

2.1 Scale Development and Item Writing

Self-report scales are widely used in the social sciences to measure beliefs, attitudes, opinions, and other subjective constructs Chan, 2010; Stanton, Sinar, Balzer, and Smith, 2002. As such they are distinct in purpose from objective tests of knowledge or reasoning, where each item typically has a single correct answer. Many self-report scales use a composite score computed from a series of items—where the respondent indicates a response to each item on a common, multi-step scale—in an effort to position each respondent on some continuum that reflects the construct definition.

Self-report scale development is often depicted as a multistep process that begins with development of a construct definition. Several publications describe methods, principles, and advice for scale development (Morgado, Meireles, Neves, Amaral, & Ferreira, 2017). Item writing typically serves as a second step in the scale construction process, after researchers have refined a construct definition sufficient to support sampling of item content from the construct domain.

Articles and chapters by Hinkin, 1995, 1998; Hinkin, Tracey, and Enz, 1997; Hinkin, 2005 emphasised the importance of writing a large enough pool of candidate items to support later deletion of poorly functioning items based on analytical results. Articles by Clark and Watson, 1995, 2019; Watson, Clark, and Tellegen, 1988 provided guidance on the item writing process by describing linguistic principles such as the avoidance of double-barrelled item wordings. Table 1 lists a few of the commonly-cited resources in this genre and includes paraphrased advice and suggestions from these authors. This advice spans different stages of the scale development process depending upon the focus of each article.

Most of the advice in Table 1 is sensible and intuitive, although it tends toward high-level principles rather than actionable rules. Even when following this advice carefully, researchers are always advised to generate a pool of candidate items larger than the expected length of the scale, with the expectation that several poorly functioning items will "wash out" of the process based on psychometric analysis of pilot data obtained from research participants.

A second commonality in the advice of authors represented in Table 1 pertains to the diversity of linguistic content. Hinkin, 1995 and others emphasised the importance of sampling widely from the content domain of the construct—a practice that can lead to substantial semantic diversity in the ideas embodied in the items. Yet achieving a satisfactory level of internal consistency with highly diverse item content also requires a scale of considerable length. As a historical example, Zelin, Adler, and Myerson, 1972 reported development of a scale for self-reported aggression that contained 64 items to measure six facets of this construct. Indeed, Schweizer, 2011 suggested that prior to

Table 1

Advice from Reviews of Item/Scale Development Practices

Author (Year)	Advice
Hinkin (1995, 1998)	Establish clear links between item wordings and the theoretical domain; Develop enough items to allow for later deletion of poorly functioning items; Use confirmatory factor analysis to document factor structure of item responses; Include a sufficiently large group of items to ensure sufficient domain sampling; Use multiple methods and samples to document construct validity
Gehlbach (2011)	Avoid reverse-scored items; Use at least five response anchors; Avoid agree-disagree response anchors; Use verbal rather than numeric labels as response anchors; Ensure that all item content is sensible for all respondents
Clark (1995)	Use simple, straightforward language; Avoid complex items, particularly double-barreled constructions; Ensure that affect-laden language does not cross-contaminate with broad individual differences (e.g., neuroticism); Avoid checklists, forced-choice formats, and visual response scales
MacKenzie (2011)	Item content must cover all sub-dimensions of construct domain; Wording should be as simple and precise as possible; Double-barreled items should be split into single ideas; Items with ambiguous terms require clarification; Complex syntax should be simplified; Items with obvious social desirability should be removed

the early 2000s, the norms for scale length tended to be in the range of 10–12 items per facet or sub-scale.

In the present day, however, a typical scale often has fewer than half that number of items. The problem of diminishing survey response rates (e.g., Fan & Yan, 2010) may now induce scale developers to limit the number of items of newly developed scales. In fact, several published articles have provided advice and methods for reducing the number of items in existing scales (Fisher, Matthews, & Gibbons, 2016; Heggestad et al., 2019; Stanton et al., 2002; Wieland, Durach, Kembro, & Treiblmaier, 2017). Smith and Stanton, 1998 suggested that with these short scales (e.g., three to five items), achieving acceptable internal consistency would require broadly worded items with overlapping semantics. Those who write survey items as part of a scale development effort have only their intuition in devising wordings that give an item a broad or narrow wording and a scale sufficient semantic diversity to adequately represent the construct.

2.2 The Pursuit of Internal Consistency

In validation studies, scale developers typically present one or more measures of reliability computed from pilot data. In studies where a scale is administered at just one point in time, a Cronbach's alpha value is often the only measure of reliability presented Cortina, 1993. Historically, alpha was created as an improved replacement for split-half reliability and is often referred to as the extent of inter-relatedness among a set of items (Dunn, Baguley, & Brunsden, 2014; Sijtsma, 2009). Alpha has been criticised on a variety of fronts: it favours scales with a larger number of items; it fluctuates when computed from different samples; it cannot not confirm the unidimensionality of a scale; the conventional minimum threshold of 0.70 is arbitrary; and researcher efforts to maximise alpha tend to homogenise item content (Cho & Kim, 2015).

From a technical standpoint, the assumptions underlying alpha are rarely met in practice (Dunn et al., 2014). For theoretically accurate usage, alpha must be computed on item data that fit an "essential tau-equivalence" model where item covariances are identical across the board. In practice, the item data from most self-report scales instead only fit a "congeneric" model (Graham, 2006), which relaxes this key assumption of essential tau-equivalence. When alpha is computed over item data that fit a congeneric model but not an essential tau-equivalence model, the results inaccurately estimate the actual reliability of a scale.

Despite the fact that the literature has offered usable alternatives to alpha (e.g., omega, see Dunn et al., 2014), scale developers do not frequently use them, perhaps in part because of limited exposure to alternative reliability measures during graduate methods training (Aiken, West, & Millsap, 2008; Oswald, Wu, & Courey, 2022) and in part because peer reviewers have come to expect a report of alpha for every multi-item scale. As a result, and despite the identified flaws of alpha, it is effectively the only universal quantity that can be extracted from validation reports to represent the reliability of a scale under development.

One of the prominent critics of alpha, Sijtsma, 2009, has clarified his position on the use of the metric. Using a set of proofs and a Monte Carlo study, this work shows that from a practical standpoint, alpha is still a useful reliability metric for scale developers and users to compute and report despite its flaws (Sijtsma & Pfadt, 2021). When analysing congeneric item data, alpha does represent a lower bound for reliability, but the discrepancy is negligible assuming four conditions are met. First, items must all be assessed on the same multi-point scale (e.g., a five-point Likert scale). Second, the scale must be effectively unidimensional, a status that scale developers are expected to establish using factor analysis. Third, the sample size used to compute psychometric data must be substantial: Sijtsma and Pfadt, 2021 showed that as sample size grows from $n = 100$ to $n = 500$, the underestimation discrepancy tends to become negligible. Finally, the items and scale under examination must fit a reflective measurement model—i.e., the items are expected to have substantial correlations with one another that reflect a common cause.

2.3 Re-Purposing Items and Scales After Validation

Research methods textbooks frequently mention the importance of reusing published, validated scales instead of fielding untested items. In the *Handbook of Research Methods in Industrial and Organizational Psychology* Rogelberg and Brooks-Laber, 2002, pg. 480 suggest that high-quality measurement is the number one challenge in the advancement of research: "... without good measures, we cannot effectively tackle research questions and advance as a science." Journal editors also repeatedly emphasise the importance of reusing thoroughly validated measures (Kuckertz, 2017; Vandenberg, 2007). While this goal may occasionally be impeded by commercialisation interests (see Creswell & Creswell, 1994, pg. 121), researchers now routinely publish the texts of their items alongside validation data. For example, Orosz, Tóth-Király, and Bőthe, 2016 published a validation article for a brief, four facet scale measuring the intensity of self-reported social media usage that contains a complete list of the item texts. Given the importance of citations to the success of an article—and by extension the careers of the authors—inclusion of a complete set of item texts in a validation article probably encourages reuse of a scale by other researchers.

This latter point raises the possibility that the linguistic content of a scale may connect with the citation frequency of a validation article. Bibliometric studies make evident that over the long haul—and after taking into account disciplinary differences in publication practices—a publication that garners more citations offers broader impacts and greater scientific influence than one with fewer citations (Chew & Relyea-Chew, 1988; Lawani, 1986; Wallin, 2005). The previous statement should not be interpreted as suggesting that the quality of a study is adequately reflected by its citation rate: specialised areas of study naturally have lower citation rates based on the smaller size of the research community. Another note of caution is also important: Remembering Dunnette's (1966) warnings about fads and fashions in social and behavioural sciences, the citation rate of a validation article may have as much to do with the popularity of the construct in question as it does with the quality of measurement. Even after taking these caveats into account, however, a scale validation article that receives many citations implies a set of items that, all else being equal, has been frequently chosen for reuse by other researchers.

In the same vein, a scale with satisfactory internal consistency and unidimensional factor structure is perhaps more likely to succeed as a predictor or outcome variable in future research studies. Examples in this genre include scales like the "mini-IPIP" (Donnellan, Oswald, Baird, & Lucas, 2006) which has been cited more than 400 times per year since publication, as well as the Positive Affect-Negative Affect Scale (PANAS Watson et al., 1988), which has been cited more than 1300 times per year since publication. These and many other examples of popular scales suggest that a well-constructed set of items that captures a construct of broad interest and that is published in a reputable journal will encourage reuse by researchers. Only a small part of that success can likely be attributed to the wording of items, but perhaps enough to warrant investigation. In the next section of the literature review, we explore methods of using natural language processing to assess the linguistic connections among items.

2.4 Algorithmic Summarization of Text

Natural language processing (NLP) is an area of research at the intersection between computer science and linguistics focusing on the development of techniques by which computers can exhibit human-like success at analysing and/or generating written text. The area has notched many successes: the modern search engine is an exemplar of many years of NLP research. In recent years, NLP has reached a level of sophistication sufficient to contribute to a variety of social science measurement tasks. Some of these tasks take the form of translating free-form text directly into a measurement, such as a personality score (Speer, 2018). NLP techniques also contribute to methods for scoring sentiment such as those found in customer comments (e.g. Ordenes, Theodoulidis, Burton, Gruber, & Zaki, 2014) or patient narratives (Provoost, Ruwaard, van Breda, Riper, & Bosse, 2019). NLP has been used to compute scores from human-authored texts that are in turn used as outcomes or predictors in a later analysis (Hellesø, 2006; Sayeed et al., 2011).

Only a few articles have applied NLP to scale development tasks (Agogo & Hess, 2018; Hernandez & Nie, 2022; Salminen, Chhirang, Jung, & Jansen, 2021). Both Agogo and Hess, 2018 and Salminen et al., 2021 focused on using NLP to harvest prospective item content from sources like social media. Hernandez and Nie, 2022 used a "chat" application to assist in developing an initial item pool.

In light of the item creation advice exemplified by the articles cited in Table 1, we know that the vocabulary, grammar, and semantics of item texts influence the characteristics of the resulting scale. The advice to write more items than will be needed in the final scale embeds the assumption that some items will function poorly when analysed. Often, we can only speculate about why a promising item performs poorly. Recent developments in machine learning, however, have enabled computational linguistic analysis of short texts that is readily available to all researchers.

Starting in the 1990s, techniques such as latent semantic analysis (LSA Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) and latent Dirichlet allocation (LDA Blei, Ng, & Jordan, 2003) opened the door to numeric representation of the meaning of a text (semantics). These representations took the form of a string of real numbers obtained by analysing the distribution of terms (words) across documents (which could be as small as a single phrase or sentence). More recently, high-dimensional numeric representations of words and sentences have benefited from progress in the use of "deep learning" neural network models. Notably, Mikolov, Sutskever, Chen, Corrado, and Dean, 2013 used deep learning on a large body of text documents to develop numeric representations known as word embeddings. A word embedding equates a word such as "company" with a unique list of dozens or hundreds of positive and negative real numbers—often referred to as a "vector." Each element of a vector is believed to encode some aspect of the semantics of the word. Mikolov et al., 2013 and many subsequent research articles have demonstrated that words with similar meanings have similar vector representations (see Bojanowski, Grave, Joulin, & Mikolov, 2017; Chandrasekaran & Mago, 2021; Kenter & De Rijke, 2015).

Refinements to this initial approach to word embedding have provided greater sophistication both in the creation and use of these vectors. Consider for example, that the word "company" has at least two senses—a formal organ-

isation or alternatively a group of people who come over for dinner. The word vector for company would necessarily represent a compromise between those different word senses. Rather than settle for this compromise, however, newer methods process words in their context, for example by creating a vector to represent a phrase, complete sentence, or whole paragraph. Reaching this level of sophistication requires substantial amounts of training data and computational resources. As a shortcut to facilitate wider use of these "sentence vectors," researchers have created and stored pretrained models that can quickly and easily be reloaded for reuse (Han et al., 2021). Research results show that these pretrained models can achieve human-like success in difficult tasks like foreign language translation (Popel et al., 2020).

We collected the item texts from a large set of validation articles and used a pretrained sentence embedding model to generate vector representations for each item in each scale. Then, for each scale, we computed a matrix of similarity values—conceptually similar to a correlation matrix—representing the semantic correspondences among each pair of items in a given scale. We surmised that, just as a correlation matrix of item responses contains valuable psychometric information, a matrix of similarity values might contain patterns that could shed light on how item texts may contain semantic relations of interest. In the next section, we describe a strategy for creating predictive models using these similarity matrices.

2.5 Convolutional Neural Networks

Convolutional neural networks (CNNs) represent a common type of computational model used for tasks such as image recognition (O'Shea & Nash, 2015). Convolution is a term borrowed from calculus that refers to the integral of the product of two functions. A CNN model uses this idea to process complex data inputs, such as the two-dimensional matrix of pixels found in a digital photograph, by using a sliding window that processes small neighbourhoods of points (Albawi, Mohammed, & Al-Zawi, 2017). After a CNN model is trained on some data, it can extract important features from these neighbourhoods and combine them into high-level summaries. As an example, image recognition of a photo of a cat might start at the level of noting characteristic shapes, gradually aggregating an area of the picture such as an ear, and finally integrating the presence of several elements such as ears, eyes, nose, and tail into a prediction that the image being analysed represents a cat. CNN models can thus be trained to recognise and make use of patterns in any dimensional matrix of data.

The CNN model training process uses training data to adjust coefficients in a model—referred to as weights and biases—in a way that gradually minimises an error function such as mean squared error. Starting in 2015, CNNs began to exceed human accuracy in a variety of predictive tasks (Ajit, Acharya, & Samanta, 2020). CNNs have since found a great variety of applications in natural language processing (Goldberg, 2017) because the architecture is good at processing localised dependencies in sequential data—a typical characteristic of natural language text. Because a CNN model can "see" neighbouring data points at the same time, it can recognise patterns that have predictive value. With appropriate data preparation, a CNN model can use variably-sized two-dimensional matrices as predictor data along with a continuously valued criterion value such as Cronbach's alpha.

2.6 Research Hypotheses

We used a pretrained sentence summarisation model to create numeric representations of the semantic content of item texts. For the set of items in each scale, we then used these numeric representations to compute a two-dimensional symmetric matrix of semantic similarity values. The resulting collection of matrices became the "predictors" in our CNN models. CNN models are well suited to use twodimensional matrices as input, because of their capability to slide a receptive window over a matrix to extract localised patterns of interest. In a manner similar to how a human analyst might scan a correlation matrix to look for notable patterns of correlations among groups of items, this strategy allowed us to use item similarity matrices as input to a CNN model whose task was then to predict an outcome variable. Based on the forgoing literature review of the item and scale development process, we pursued three research hypotheses:

- *Hypothesis 1*: Item similarity matrices will predict the Cronbach's internal consistency reliability of multiitem scales.
- *Hypothesis 2*: Item similarity matrices will predict the citation rate of the validation article in which the items and the scale validation evidence appear.
- *Hypothesis 3*: Item similarity matrices will predict the impact factor of the journal where the validation article for the multi-item scale was published.

While the three criteria mentioned in these hypotheses only imperfectly represent the "quality" of a measurement scale, any meaningful capability to predict one or more of these criteria would suggest that the linguistic relations among items may have systematic, detectable connections to the qualities of a scale. If that idea is supported, the NLP tools for computing numeric representations of items and for computing linguistic similarity among items may

eventually have useful application as an additional tool in the scale development process.

3 Method

Our research team searched the social science literature for journal articles that reported on self-report "scale development" and "validation," using those search terms in Google Scholar and other research databases to yield a list of many thousands of potential articles. We evaluated each candidate article to make sure that full item texts were shown; that psychometric data including Cronbach's alpha were provided; that the items in question represented a unidimensional scale; and that the article appeared in a journal for which a three-year impact factor value was reported in 2021. The dataset mainly comprises citations to journals in applied psychology, allied health fields, and business. All validation articles included in the sample were for scales/items written in English that were administered to English-speaking research participants.

For the analysis described below, we extracted item texts and other data from $n = 386$ scale development and validation articles (full reference list available upon request). An example was Heatherton and Polivy (1991), who developed a multi-item, self-report scale for measuring state self-esteem. For each article, we extracted the texts of the items for each scale. Note that we stored the item texts in the order of appearance in the article so that the CNN model might "see" them in the same order in which participants would typically read them. It is important to note that when items are validated or reused, they are sometimes presented in a different order from how the items are displayed in the validation article and, likewise, that items for one scale are sometimes interspersed with items from another scale.

For each scale, we recorded the Cronbach's alpha internal consistency reliability as reported in the publication. As alpha only assesses an "internal" quality of a scale, we also wanted to have metrics that captured some other qualities. Thus, we also recorded the log of the number of citations per year that the article had received at the time we retrieved it using values displayed in Google Scholar. We interpreted this figure as a proxy for the popularity of the scale among researchers. We used the log transformation on citations per year because the distribution of the raw metric was highly positively skewed. Finally, we used the log of the three-year journal impact factor from 2021 as a proxy for the quality of the journal where the validation article was published. Again here, the log transformation was used because the distribution of impact factors was highly positively skewed.

We converted the text of each item in each scale to a high dimensional vector representation using bidirectional encoder representations from transformers, commonly known as BERT, to create each numeric representation. The pre-trained model we used is known as "all-MiniLM-L6-v2" and was derived by "fine-tuning" the Wang et al., 2020 MiniLM model on 1.7 billion English sentences from a wide variety of sources. The method used to create all-MiniLM-L6-v2 was optimised to enhance a measure of dissimilarity between unlike sentences. Note that in this context, fine tuning means that an initial model such as the one provided by Wang et al., 2020 was subjected to additional training to improve the model's suitability for a particular task—in this case sentence summarization.

When we submitted the text of an item to all-MiniLML6-v2, it returned a 384-dimensional floating point numeric vector. The values in a vector can be loosely conceptualised as a set of weights representing the item's meaning in 384 different, unspecified semantic dimensions. As a result of how all-MiniLM-L6-v2 was trained and fine-tuned, semantically similar sentences have vector representations placing them close to one another in 384-dimensional space. Note that we also experimented with higher dimensional models, such as the 768-dimensional "msmarco-distilbert-base-v3" and found that using these alternative models had negligible effects on the results of our downstream analyses. This finding was consistent with Yin and Shen, 2018, who found diminishing returns in model performance with increasing dimensionality.

We used a measure known as cosine similarity, commonly used in text processing applications, to calculate a matrix of semantic similarity values among the items for each scale. Note that there are several ways of calculating and normalising cosine similarity and we used a method that captures similarity on a scale ranging from 0 (most distant) to 1 (most similar). As an example, imagine three Likert-type statements where the first two are quite similar to each other (cosine distance near 1) and the third item refers to something different (cosine distances near 0). These three items would be represented by a 3×3 symmetric square matrix of values with ones on the diagonal, conceptually similar to the contents of a correlation matrix. The median length of a scale in our data set was five items, so a typical matrix thus contained 25 cells, with ones on the diagonal and the upper and lower triangles containing identical information.

4 Results

Prior to training the CNN models, we conducted linear regression analyses to predict each of the three outcome variables. Regression models cannot handle predictor data comprising matrices of varying size, so we summarised the similarity matrix for each scale by computing the mean and the standard deviation of the similarities, using the lower

Table 2

Linear Regression Analysis Results

Dep. Var	Mean Sim	SD Sim	Num. Items	Adj. R-squared
Cronbach's Alpha	–0.85*	–0.85*	0.00	0.09
Log Citations per Year	0.33	–0.11	–0.01	0.01
Log Impact Factor	0.08	1.35	0.01	0.01

* Indicates a significant regression weight. Regression weights are unstandardised. Only the regression on Cronbach's alpha was statistically significant at $p < 0.05$.

triangle from each similarity matrix. We also used the number of items in each scale as a control variable. These analyses served two closely related goals: First, the analyses would indicate a baseline level of predictive success that each CNN model would need to surpass to be useful. Second, the analyses would show what could be achieved with a simplistic approach to summarising inter-item similarity and the amount of variation in that metric. Table 2 shows the results.

Results suggest that the mean similarity and the standard deviation of similarity each predict the log of Cronbach's alpha. When the mean similarity among items is larger, the Cronbach's alpha is higher. When the variability among similarity values is larger the Cronbach's alpha is lower. The adjusted R-squared of 0.09 for the prediction of the of Cronbach's alpha represents a lower bound that we hoped the corresponding CNN model would exceed. The analysis of log citations per year and log impact factor were not statistically significant.

Next, we trained three CNN models, each using the scale/article as the unit of analysis and each using the collection of similarity matrices as predictor data. We used the same simple model geometry for all three models: a single convolutional layer with a "window" that examined 3 × 3 neighbourhoods of the similarity matrix, eight trainable "kernels," and a 2 × 2 "max pooling" pattern feeding into a dense layer of 20 nodes with linear activation. Each kernel, sometimes also called a filter, provides one trainable system of weights for processing input data obtained from the window. Having eight of these creates eight unique patterns that can be learned. Max pooling reads the output of each kernel and focuses on the most notable aspect. The dense layer with 20 nodes acts like a tiny "brain" to learn the optimal way of combining the outputs of the max pooling layer. A final layer consisting of a single linear node combines the results of the dense layer into predicted values for each model.

Relative to a typical machine learning application, these models are quite small and therefore easier to train. For example, Kim, Seo, Yoo, and Shin, 2022 satisfactorily trained a CNN model containing 100 times more trainable weights than this model using a sample size of only $n = 500$. Likewise, Brigato and Iocchi, 2021 trained a simple CNN model about the same size as this one to a satisfactory level of accuracy using a sample size of $n = 320$. A more complete explanation of model configurations, variations, and hyperparameters appears in the Appendix.

As noted above, the median number of items per scale was five (min = 2, max = 34). Thus, a typical predictor data element was a square symmetric matrix with 25 entries and ones on the diagonal. Ignoring diagonals, the mean cosine similarity value across all input matrices was 0.436 on a scale of zero to one. Note that this mean value was computed by summing the individual off-diagonal elements extracted from all $n = 386$ input matrices.

Table 3 shows descriptive statistics for the dependent variables. The log of citations per year and the log of the journal impact factor were positively correlated. The correlation between alpha values and the log of citations per year was negative and statistically significant but minuscule. The correlation between alpha values and the log of journal impact factor was close to zero.

At 0.09, the standard deviation of alpha values showed that there was limited variation in the reported values of alpha in validation articles, no doubt because few scales are published with alpha values below 0.70. The mean of the log of citations per year corresponds to an average about 2.1 citations per year. The mean of the log of journal impact factor corresponds to a three-year JIF of 1.5. As noted above, log transformed data was used for these variables because both distributions of the raw metrics were highly positively skewed.

Note that in many applications of machine learning, a dataset is divided into subsections, with the goal of training the model on a majority of the data and conducting crossvalidation analysis using a held-back portion. This strategy works satisfactorily when there are thousands of cases or more. For our dataset of $n = 386$, a single randomised split into training and validation sets might easily capitalise on chance. For this reason, we used k = 10-fold cross-validation, a method that repeats the analysis multiple times, each time using a different subset of 90% of the

Table 3

Descriptive Statistics for Dependent Variables and Mean Cosine Similarity

Variable	Mean	SD	1. Alpha	2. Log CPY	3. Log JIF
1. Cronbach's Alpha	0.82	0.09	–	–	–
2. Log CPY	0.75	0.70	–0.10**	–	–
3. Log JIF	0.43	0.26	–0.01	0.24***	–

$n = 386$, $^{**}p < 0.05$, $^{***}p < 0.001$

data for training and the remaining 10% for cross-validation (Refaeilzadeh, Tang, & Liu, 2009). K-fold cross-validation has shown utility in accurately estimating prediction error, with k = 5 and k = 10 shown to be suitable across many analytic situations (Rodriguez, Perez, & Lozano, 2009). An inspection of meansquared error values for each fold showed minimal variation, suggesting that the training process was stable across folds for all three models. The mean-squared error values in Table 4 represent the mean of the final model error values across the k = 10 runs, along with the corresponding R-squared values.

Results showed success in predicting each of the three criteria, with the highest R-squared value for the log of the impact factor (0.208) and the lowest for the log of citations per year (0.164). A traditional F-test of the significance of these R-squared values is not possible from a CNN model, but the corresponding Pearson' r value for each of the R-squared values shown in Table 3 would be statistically significant at $p < 0.001$. Thus, all three hypotheses were supported. Notably, all three of the R-squared values substantially exceeded the baseline value implied by linear regression results shown in Table 2. These results clearly indicated that each of the three models keyed off of patterns of information in the similarity matrices that transcended a simple averaging of the cosine similarities.

We graphed activation patterns from the layers of each model to shed light on how the models made their predictions. This process created heat maps showing which cells from an input matrix most strongly influenced the model's predictions. We graphed these activation patterns by making a prediction using a variety of example input matrices. Fig. 1 shows a typical activation pattern from the max pooling layer of the model. It is useful to examine the max pooling layer, because this layer identifies the most prominent features emerging from the convolutional kernels. The grey scale heat maps shown in Fig. 1 use black to represent the lowest depicted activation value and white for the highest activation value. One hundred different gradations of grey capture the variation in between the lowest and highest values.

Recall that each of our CNN models had eight kernels, hence the eight panes of the display in Fig. 1. The numbering that appears along the X and Y-axes indexes the set of activation values emerging from the max pooling process applied after the previous CNN layer. Each of the eight panes of the figure thus displays a 15 × 15 matrix of the most prominent features emerging from the convolutional layer. The diagonals are strongly activated in the outputs of five of the panes and weakly in two additional panes. Because the diagonals of the input matrices always contain ones, this pattern is expected. Note that many of these panes also show activation of the sub diagonal—the cells just above and/or below the main diagonal. These positions represent the pairwise similarities of neighbouring items (i.e., item 1 to item 2, item 2 to item 3, etc.) in a given scale. This is a unique finding that was consistent across the models for all three independent variables and across a wide variety of test instances used to activate the models. Keeping in mind that the kernel weights in the preceding layer are trained to optimise the predictive capability of the model, this finding indicates that the models are recognising when each item has an above-average degree of similarity with its immediately preceding neighbour.

As a final step, we conducted a sensitivity analysis to understand whether variations in model architecture would change model performance. Models with an additional convolutional and max pooling layer required more training epochs, but did not achieve superior results. Modifying key hyperparameters of the reported model geometry within typical ranges—the learning rate, the number of epochs, the number and shape of the convolutional kernels, the size of the max pooling layer, or the size of the dense layer—did not substantially change the model results shown in Table 4.

Table 4

Convolutional Neural Network Analysis Results

Dependent Variable	Variance	Average MSE for k = 10-fold CV	Corresponding R-squared value
Cronbach's Alpha	0.009	0.007	0.188
Log Citations per Year	0.489	0.409	0.164
Log Impact Factor	0.069	0.055	0.208

MSE mean squared error

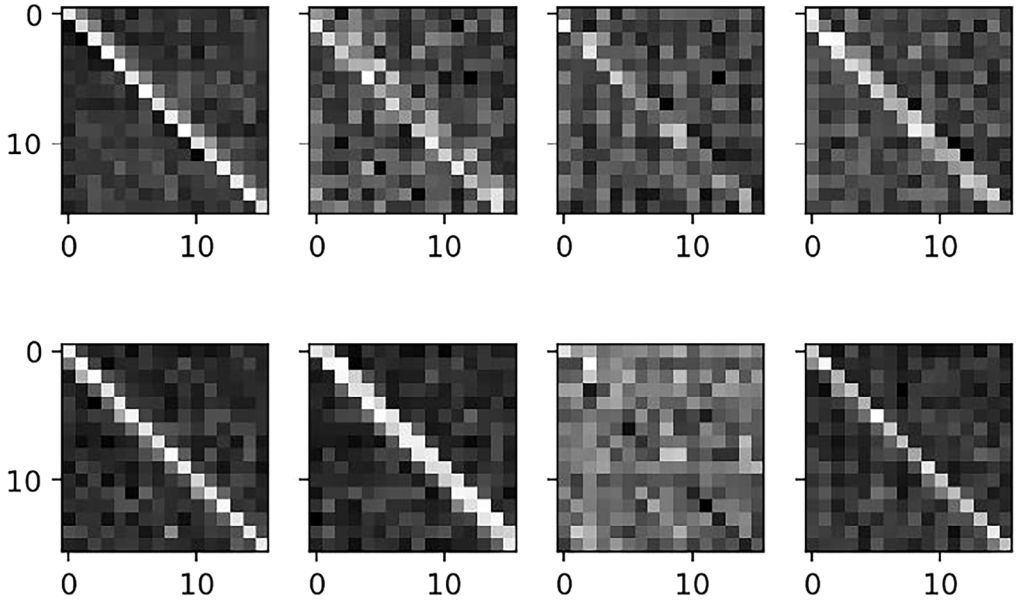

Fig. 1

Grey Scale Heat Map of Max Pooling Activation Patterns

We tested three variations of learning rate alongside variations in the number of training epochs up to 30 epochs. After settling on a learning rate of 0.001 and 30 training epochs with k = 10-fold cross-validation we then tested three variations of kernel size, three variations of max pooling size, and two variations of the size of the dense layer. The model described above and in the Appendix worked best among the variations, but none of the tested variations caused more than minimal change in the reported R-squared values.

5 Discussion

We reviewed principles of item writing and scale construction suggesting that scale developers often begin with a construct definition and then compose a large set of candidate items that address ideas fitting the construct. When research participants later provide item responses, psychometric results indicate which items should be retained and discarded. Among items that remain, the pattern of covariance (of participant responses) among them drives the Cronbach's alpha internal consistency.

Particularly for shorter scales, research has suggested that somewhat homogeneous item content is needed to achieve satisfactory internal consistency. Results of a simple linear regression analysis supported this idea. Cronbach's alpha has a significant and positive relationship to the mean similarity values in a semantic similarity matrix. Likewise, another significant finding was that the greater the amount of variation in similarity values, the lower the alpha value was. The mean similarity and standard deviation of similarity did not predict either citations per year or the journal impact factor.

In contrast, CNN models outperformed linear regression and were successful in predicting Cronbach's alpha, citations per year, and the impact factor of the journal using the patterns detected in similarity matrices as predictors. The fact that these CNN models substantially surpassed the linear regression models signifies that complex patterns of semantic similarity are being detected and connected to the outcome variables. In particular, a graphical display of activation patterns in the CNN model suggest that semantic connections between neighbouring items may be important.

By its nature, a CNN model captures localised features, amplifies the important ones, and aggregates those to predict the outcome. The 3 × 3 kernel size—a common choice in CNN modelling—is considered ideal for picking up localised patterns among nearby points (Singh et al., 2019). In the analysis we conducted, several of the eight trained kernels keyed off of elements of the matrix that represented the similarities of "next-door neighbour" items. For a scale with five items, there are four such values in the "subdiagonal," i.e., the values just below (or above) the main diagonal of the similarity matrix, suggesting that the linguistic similarity of an item to the item immediately succeeding it matters. This is consistent with a body of research showing that item sequencing impacts both item responses and the properties of a scale (Hayes, 1964; Knowles, 1988; Smyth, Israel, Newberry III, & Hull, 2019; Steinberg, 1994). Note

that this finding neither argues for nor requires a high degree of linguistic similarity among all the items in a scale, and is therefore not an argument for linguistic homogeneity among items. Rather, it suggests that respondents may respond to the order in which they encounter a set of items. If two items sampling a similar aspect of a construct appear next to one another, respondents may be more likely to provide item responses that increase alpha.

When a validation article for a scale describes satisfactory psychometric performance, that article might become eligible for publication in a more prominent journal and may gain citations reflecting its popularity of reuse. The number of citations of a validation article and the impact factors of its journal were positively correlated, an intuitively satisfying (though somewhat tautological) finding that suggests that future researchers are more likely to find, reuse, and cite a validation article from a more prominent journal than from a less prominent one. Moreover, our CNN analysis showed that patterns of linguistic similarity among items predicted both the journal impact factor and the citation rate for a scale validation article. This suggests that linguistic aspects of the items in a scale may influence researchers' choices to reuse the scale in their own work. These results also suggest the possibility that the wording of items may somehow indirectly influence the opinions of peer reviewers and editors who review and make decisions on validation articles.

We conducted a sensitivity analysis on the CNN model showing that our model configuration decisions were sound and unaffected by minor variations in the model architecture or hyperparameters. Those variations included the use of larger kernel sizes (5×5, 7×7, and 9×9) commonly used in CNN model tests. The fact that these larger filters did not improve model performance substantiates the idea that the similarity patterns among nearby items provides predictive power to the models.

6 Limitations

One limitation of the current study plagues many uses of machine learning in social science research: Unlike simple linear models, where the coefficients are readily interpreted, neural network models contain hundreds or thousands of weights whose individual roles in a model's performance are not readily discernible. We chose the CNN model because it is suitable for ingesting variably-sized, two-dimensional matrices as input data for a predictive model and because a CNN model facilitates the recognition of patterns in matrices. These benefits must be weighed against the relatively poor visibility into how the CNN model produces its results. In future research, instead of the simple visualisations we interpreted qualitatively to understand model activation patterns, researchers might develop more structured techniques for understanding and reporting model activation patterns.

Another limitation lies in the nature of our independent variables. Cronbach's alpha is a popular metric to report, but as discussed in the introduction, alpha represents a lower bound for reliability, and using it properly requires meeting a set of assumptions that are rarely met in practice. In addition, internal consistency represents just one narrow aspect of the quality of a set of items. Other aspects of scale performance—such as factor purity, test-retest reliability, and criterion related validity evidence—are more difficult to study but could also be more informative. Likewise, variations in citation rates and journal impact factors are subject to a wide variety of influences that have little to do with the quality or usefulness of a scale. In future research it would be valuable to delve into additional dependent variables that capture more fully those aspects of a scale that make it suitable for reuse in future research efforts.

Even with these limitations in mind, there are practical applications of our results: Scale developers working on the item writing phase of scale development can easily create sentence embeddings for a set of proposed items and compute cosine similarities among them with just a few lines of Python code. Researchers could select the presentation order of a set of items based on putting semantically similar items next to each other. In addition, consider a situation when a researcher reuses a previously validated item but with minor modifications to the original wording. A modified wording for an item could easily be compared to the original and revised to improve the similarity. The present research also lays the groundwork for making the guidelines for item writing more concrete and specific such that the wording of entries in an initial item pool could be more thoroughly refined before the first pilot data set is collected. The Appendix contains Python code for computing cosine similarity among phrases or sentences to enable further experimentation.

7 Conclusion

The work described here suggests that potential exists for applying contemporary natural language processing and machine learning techniques to scale development. The present study provides systematic, quantitative evidence that the wording of item texts as presented in validation articles connects with some of the resultant qualities of a scale. Further, the results suggest that one way of achieving good results may arise from attending to the semantic linkages between neighbouring items. More generally, semantic summarisation using pretrained NLP models holds

promise as an additional tool that researchers can use when developing a new self-report scale.

References

Agogo, D., & Hess, T.J. (2018). Scale development using twitter data: applying contemporary natural language processing methods in is research. In *Analytics and data science: advances in research and pedagogy* (pp. 163–178).

Aiken, L.S., West, S.G., & Millsap, R.E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: replication and extension of aiken, west, sechrest, and reno's (1990) survey of phd programs in north america. *American Psychologist*, *63*(1), 32.

Ajit, A., Acharya, K., & Samanta, A. (2020). A review of convolutional neural networks. In *2020 international conference on emerging trends in information technology and engineering (ic-etite)* (pp. 1–5). IEEE.

Albawi, S., Mohammed, T.A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. In *2017 International conference on engineering and technology (icet)* (pp. 1–6). IEEE.

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*, 993–1022.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Brigato, L., & Iocchi, L. (2021). A close look at deep learning with small data. In *2020 25th international conference on pattern recognition (icpr)* (pp. 2490–2497). IEEE.

Calderón, J.L., Morales, L.S., Liu, H., & Hays, R.D. (2006). Variation in the readability of items within surveys. *American journal of medical quality*, *21*(1), 49–56.

Chan, D. (2010). So why ask me? are self-report data really that bad? In *Statistical and methodological myths and urban legends* (pp. 329–356). Routledge.

Chandrasekaran, D., & Mago, V. (2021). Evolution of semantic similarity—a survey. *ACM Computing Surveys (CSUR)*, *54*(2), 1–37.

Chew, F.S., & Relyea-Chew, A. (1988). How research becomes knowledge in radiology: an analysis of citations to published papers. *American Journal of Roentgenology*, *150*(1), 31–37.

Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: well known but poorly understood. *Organizational research methods*, *18*(2), 207–230.

Clark, L., & Watson, D. (1995). Constructing validity: basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309–319.

Clark, L., & Watson, D. (2019). Constructing validity: new developments in creating objective measuring instruments. *Psychological assessment*, *31*(12), 1412–1444.

Cortina, J.M. (1993). What is coefficient alpha? an examination of theory and applications. *Journal of applied psychology*, *78*(1), 98.

Creswell, J.W., & Creswell, J.D. (1994). *Research design: qualitative and quantitative approaches*. SAGE.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391.

Donnellan, M.B., Oswald, F.L., Baird, B.M., & Lucas, R.E. (2006). The mini-ipip scales: tiny-yet-effective measures of the big five factors of personality. *Psychological assessment*, *18*(2), 192.

Dunn, T.J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: a practical solution to the pervasive problem of internal consistency estimation. *British journal of psychology*, *105*(3), 399–412.

Dunnette, M.D. (1966). Fads, fashions, and folderol in psychology. *American psychologist*, *21*(4), 343.

Edmondson, D. (2005). Likert scales: a history. In *Proceedings of the conference on historical analysis and research in marketing* (Vol. 12, pp. 127–133).

Fan, W., & Yan, Z. (2010). Factors affecting response rates of the web survey: A systematic review. *Computers in human behavior*, *26*(2), 132–139.

Fisher, G.G., Matthews, R.A., & Gibbons, A.M. (2016). Developing and investigating the use of single-item measures in organizational research. *Journal of occupational health psychology*, *21*(1), 3.

Gehlbach, H., & Brinkworth, M.E. (2011). Measure twice, cut down error: A process for enhancing the validity of survey scales. *Review of general psychology*, *15*(4), 380–387.

Goldberg, Y. (2017). *Neural network methods for natural language processing*. Morgan & Claypool.

Graham, J.M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: what they are and how to use them. *Educational and psychological measurement*, *66*(6), 930–944.

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., & Zhu, J. (2021). Pre-trained models: Past, present and future. *AI Open*, *2*, 225–250.

Hayes, D.P. (1964). Item order and guttman scales. *American Journal of Sociology*, *70*(1), 51–58.

Heatherton, T.F., & Polivy, J. (1991). Development and validation of a scale for measuring state self-esteem.

Journal of Personality and Social psychology, *60*(6), 895.

Heggestad, E.D., Scheaf, D.J., Banks, G.C., Hausfeld, M.M., Tonidandel, S., & Williams, E.B. (2019). Scale adaptation in organizational science research: A review and best-practice recommendations. *Journal of Management*, *45*(6), 2596–2627.

Hellesø, R. (2006). Information handling in the nursing discharge note. *Journal of Clinical Nursing*, *15*(1), 1121.

Hernandez, I., & Nie, W. (2022). The ai-ip: Minimizing the guesswork of personality scale item development through artificial intelligence. *Personnel Psychology*.

Hinkin, T.R. (1995). A review of scale development practices in the study of organizations. *Journal of Management*, *21*(5), 967–988.

Hinkin, T.R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods*, *1*(1), 104121.

Hinkin, T.R. (2005). Scale development principles and practices. In R.A. Swanson & E.F. Holton III (Eds.), *Research in organizations: Foundations and methods of inquiry* (pp. 161–179). San Francisco: Berrett-Koehler.

Hinkin, T.R., Tracey, J.B., & Enz, C.A. (1997). Scale construction: developing reliable and valid measurement instruments. *Journal of Hospitality Tourism Research*, *21*(1), 100–120.

Jebb, A.T., Ng, V., & Tay, L. (2021). A review of key likert scale development advances: 1995–2019. *Frontiers in Psychology*, *12*, 1590.

Kenter, T., & De Rijke, M. (2015). Short text similarity with word embeddings. In *Proceedings of the 24th acm international on conference on information and knowledge management* (pp. 1411–1420).

Kim, D., Seo, S.B., Yoo, N.H., & Shin, G. (2022). A study on sample size sensitivity of factory manufacturing dataset for cnn-based defective product classification. *Computation*, *10*(8), 142.

Knowles, E.S. (1988). Item context effects on personality scales: measuring changes the measure. *Journal of Personality and Social Psychology*, *55*(2), 312.

Kobayashi, V.B., Mol, S.T., Berkers, H.A., Kismihók, G., & Den Hartog, D.N. (2018). Text mining in organizational research. *Organizational Research Methods*, *21*(3), 733–765.

Kuckertz, A. (2017). Measuring entrepreneurship—a collection of valid scales. *International Journal of Entrepreneurial Behavior & Research*, *23*(1), 56–58.

Lawani, S. (1986). Some bibliometric correlates of quality in scientific research. *Scientometrics*, *9*(1-2), 13–25.

MacKenzie, S.B., Podsakoff, P.M., & Podsakoff, N.P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS quarterly*, *35*(2), 293–334.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality* (pp. 3111–3119).

Morgado, F.F., Meireles, J.F., Neves, C.M., Amaral, A., & Ferreira, M.E. (2017). Scale development: ten main limitations and recommendations to improve future research practices. *Psicologia: Reflex ao e Crítica*, *30*.

Ordenes, F.V., Theodoulidis, B., Burton, J., Gruber, T., & Zaki, M. (2014). Analyzing customer experience feedback using text mining: a linguistics-based approach. *Journal of Service Research*, *17*(3), 278–295.

Orosz, G., Tóth-Király, I., & Bőthe, B. (2016). Four facets of facebook intensity-the development of the multidimensional facebook intensity scale. *Personality and Individual Differences*, *100*, 95–104.

O'Shea, K., & Nash, R. (2015). *An introduction to convolutional neural networks*. arXiv preprint arXiv:1511.08458.

Oswald, F.L., Wu, F.Y., & Courey, K.A. (2022). Training (and retraining) in data, methods, and theory in the organizational sciences. In *Data, methods and theory in the organizational sciences* (pp. 294–316). Routledge.

Popel, M., Tomkova, M., Tomek, J., Kaiser, Ł., Uszkoreit, J., Bojar, O., & Žabokrtsk, Z. (2020). Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, *11*(1), 4381.

Provoost, S., Ruwaard, J., van Breda, W., Riper, H., & Bosse, T. (2019). Validating automated sentiment analysis of online cognitive behavioral therapy patient texts: an exploratory study. *Frontiers in Psychology*, *10*, 1065.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Crossvalidation. *Encyclopedia of database systems*, *5*, 532–538.

Rodriguez, J.D., Perez, A., & Lozano, J.A. (2009). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence*, *32*(3), 569–575.

Rogelberg, S.G., & Brooks-Laber, M.E. (2002). Securing our collective future: Challenges facing those designing and doing research in industrial and organizational psychology. In *Handbook of research methods in industrial and organizational psychology* (pp. 479–485). Malden: Blackwell.

Salminen, J., Chhirang, A., Jung, S., & Jansen, B.J. (2021). Helping professionals select persona interview questions using natural language processing.

In *Ifip conference on human-computer interaction* (pp. 280–290). Springer.

Sayeed, A., Rusk, B., Petrov, M., Nguyen, H.C., Meyer, T.J., & Weinberg, A. (2011). Crowdsourcing syntactic relatedness judgements for opinion mining in the study of information technology adoption. In *Proceedings of the 5th acl-hlt workshop on language technology for cultural heritage, social sciences, and humanities* (pp. 69–77).

Schweizer, K.E. (2011). Editorial: Some thoughts concerning the recent shift from measures with many items to measures with few items. *European Journal of Psychological Assessment*, *27*(2).

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107–120.

Sijtsma, K., & Pfadt, J.M. (2021). Part ii: On the use, the misuse, and the very limited usefulness of Cronbach's alpha: discussing lower bounds and correlated errors. *Psychometrika*, *86*(4), 843–860.

Singh, A., Saha, S., Sarkhel, R., Kundu, M., Nasipuri, M., & Das, N. (2019). *A genetic algorithm based kernel-size selection approach for a multi-column convolutional neural network*. arXiv preprint arXiv: 1912.12405.

Smith, P.C., & Stanton, J.M. (1998). Perspectives on the measurement of job attitudes: the long view. *Human Resource Management Review*, *8*(4), 367–386.

Smyth, J.D., Israel, G.D., Newberry III, M.G., & Hull, R.G. (2019). Effects of stem and response order on response patterns in satisfaction ratings. *Field methods*, *31*(3), 260–276.

Speer, A.B. (2018). Quantifying with words: an investigation of the validity of narrative-derived performance scores. *Personnel Psychology*, *71*(3), 299–333.

Stanton, J.M., Sinar, E.F., Balzer, W.K., & Smith, P.C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology*, *55*(1), 167–194.

Steinberg, L. (1994). Context and serial-order effects in personality measurement: limits on the generality of measuring changes the measure. *Journal of Personality and Social Psychology*, *66*(2), 341.

Vandenberg, R.J. (2007). Editorial. *Organizational Research Methods*, *11*(1), 6–8.

Wallin, J.A. (2005). Bibliometric methods: pitfalls and possibilities. *Basic & Clinical Pharmacology & Toxicology*, *97*(5), 261–275.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). Minilm: deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, *33*, 5776–5788.

Watson, D., Clark, L., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, *54*(6), 1063–1072.

Wieland, A., Durach, C.F., Kembro, J., & Treiblmaier, H. (2017). Statistical and judgmental criteria for scale purification. *Supply Chain Management: An International Journal*, *22*(4), 321–328.

Yin, Z., & Shen, Y. (2018). On the dimensionality of word embedding. In *Proceedings of the 32nd conference on neural information processing systems (neurips 2018)* (pp. 895–906).

Zelin, M.L., Adler, G., & Myerson, P.G. (1972). Anger self-report: an objective questionnaire for the measurement of aggression. *Journal of consulting and Clinical Psychology*, *39*(2), 340.