Scale Semantic Overlap and its Impact on Item Response Patterns

Konrad Mikalauskas

Faculty of Social and Behavioral Sciences, University of Amsterdam

7205RMTPMY: Research Master's Thesis Psychological Methodology

Hannes Rosenbusch

September 30, 2023

Abstract

Why are people's responses to items from a personality scale correlated? Psychologists generally presume that item-item relationships are explained by a latent construct causing patterns of construct-related behaviors that the personality scale then measures. However, emerging literature suggests item responses are influenced not only by latent constructs, but also by responses to *previously* answered items due to high semantic similarity between items. This semantic overlap can coerce people to respond consistently to similar items, calling into question whether the empirical item-item relationships (i.e., response patterns) in a given scale are a property of the measured people, or the scale itself. Hence, this project investigates semantic overlap within psychological scales and its influence on response patterns. We systematically varied the amount of within-scale semantic overlap in five psychological scales by paraphrasing them using GPT-4. We then predicted the empirical item-item correlations in the original and paraphrased scales using the respective item-pairs' semantic similarities. We successfully increased semantic distance between scale items, and found a positive association between items' empirical and semantic relationships; however, we found no evidence that decreasing semantic overlap results in weaker item response patterns. These findings are discussed in the context of scale development and quantifying semantic overlap.

Scale Semantic Overlap and its Impact on Item Response Patterns

In psychology, it is generally assumed that behavior is a joint function of persons and situations (Bowers, 1973; Meyer et al., 2010; Mischel, 1973; Murphy, 1996; Wood et al., 2021). Just as persons can differ in terms of construct strength (e.g., some people are more extraverted than others), situations can also differ in their strength – i.e., in the saliency of external cues which indicate what behaviors are desirable (Meyer et al., 2010). This means that in "weak" situations, individual differences dominate the behavior-generating function, while in "strong" situations, the situations themselves best explain behavior. Consider how differences in extraversion would be expressed in the situations of a dinner with friends versus the military – people will likely have more similar patterns of extraversion-related behaviors in the military, because the situation heavily dictates what is appropriate. In their semantic theory of survey response (STSR), Arnulf and Larsen (2021) propose that language itself is a "strong" situation; that is, the language used in scale items. While people's response to the first scale item is "genuine" in the sense that it arises from reflecting on their past attitudes/behaviors; subsequent responses are increasingly dominated by the items' semantic overlap (i.e., shared item meaning) with previously answered ones. Meaning that instead of *solely* probing the behaviors, thoughts, and feelings that psychological constructs manifest as, psychological scales *also* measure the semantic similarity between items, inflating the observed inter-item correlations.

Scales as dart games

As an analogy, consider a novice darts player; for the first attempt, she stand in front of the board and throws the first dart. For the second throw though, she now also reflects on the first attempt – "last throw was too low", "this dart feels heavier", "what if I lean forward more?" – and throws again. Importantly, the two darts are not identical, nor are any of the subsequent darts she will throw (i.e., they all differ in their weight, length, shape, material, etc.). In this scenario, one can think of the darts as items, the player's throws as responses, and any specific combination of darts as a scale. Back to her throws – the *first* throw reflects *only* her pre-existing dart-throwing ability; however, each *following* throw is more and more influenced by previous ones with the goal of being accurate. In other words, subsequent throws are increasingly dominated by the overlap in physical conditions with previous throws. To observe the influence of pure skill, we would have to brainwash the player after each throw so that she reflects only on the present physical conditions. If we do not, then we are also measuring the physical relationships between the darts and how well she understands them. There is a problem of *dependence* between the darts when used as a measure of dart-throwing ability. Throwing darts is unlike responding to surveys in one important aspect though – there is no "metric" that we try to achieve when filling out surveys (like higher accuracy when playing darts). However, there is still a goal that we strive towards and that is self-consistency. People have a strong intrinsic drive to act consistently with our previous actions, regardless of the action's purpose or if we are being observed (Coombs, 1964; Michell, 1994; Sadler & Woody, 2003). This includes previous item responses, where our ability to judge two *responses* as coherent depends on how similar in meaning are the two respective *items*. Relating this to prior research, it has been shown that administering scales that have been constructed to have uninterpretable items (i.e., replacing the word 'intelligence' with a nonsense word in an intelligence attitude scale; Maul, 2017) can still result in excellent internal consistency between item responses. Moreover, for multi-facet scales, presenting items from the same facets together (e.g., A1-2-3-4, B1-2-3, etc., where 'A' and 'B' are construct facets) results in greater empirical support for the theoretical construct structure than random presentation (Şahin, 2021). These finding show that both semantic and syntactic processing of item relationships has an undeniable influence on people's response patterns (i.e., empirical inter-item correlations) that is unexplainable if assuming that psychological scales measure *only* differences in construct strength. This is not to say that semantic relationships between items determine the entire possible survey response space, but that part of the observed response patterns are contaminated by semantic processing of item relationships. Rather than being "pure" measures of construct strength with random measurement error, many psychological scales could be suffering from inflated inter-item correlations that arise due to high semantic overlap between the items (and people's desire for self-consistency). Thus, we ask how semantic overlap between items in psychological scales impacts people's item response patterns. In other words, do changes in the semantic relationships between items result in changes in the observed response patterns?

Quantifying semantic overlap

To investigate the above question, a method of quantifying semantic overlap between statements is needed. Word embeddings, a common technique in natural language processing, represent words as unique points in high-dimensional numerical space (Chowdhary, 2020; Jurafsky & Martin, 2000), and their distance from each other in this space can be interpreted as their similarity. While there are various methods of computing word embeddings, they are all based on the assumption that co-occurring words in similar contexts are semantically alike (Joos, 1950). Modern techniques employ neural networks to predict word probabilities in context, resulting in more similar words having closer vectors. Word similarity can then be quantified by calculating the distance between the respective vectors. Scale items of course carry more meaning than the sum of the individual words used to construct them. Therefore, to account for sentence-level context and semantics, we can use a model like Google's Universal Sentence Encoder (USE; Cer et al., 2018) which computes context-aware embeddings through the use of transformer architecture (Vaswani et al., 2017). Altogether, semantic overlap between two scale items can be computed by computing the distance between their sentence embeddings.

Semantic and empirical item relationships

Now that we can quantify semantic overlap between items, how do we study the link between semantic and empirical inter-item relationships? Prior research into semantic content of organizational behavior scales has demonstrated that a large percentage of the variance in people's item response patterns (60-86%; Arnulf et al., 2014) and relationships between items from different scales (25-69%; Nimon et al., 2016) can be predicted directly from items' semantic similarities. In both papers, the researchers calculated inter-item semantic similarity using latent semantic analysis (LSA) – an older statistical method for word meaning representation. They then used the computed semantic similarities between items to predict the *empirical correlations between responses*. While these results are potentially alarming for the measurement of psychological constructs, they also highlight that by manipulating the semantic overlap between items, we can potentially influence people's response patterns (and thus the interpretations of survey results).

Administering several versions of the “same” scale with differing item wordings is not novel – it is standard practice for assessing scales’ alternative-form reliability (AFR; Cook & Beckman, 2006; McNemar, 1955). We borrow thought from AFR literature and argue that if consistent inter-item empirical relationships are observed across all reasonable item paraphrases, then we can argue that the results are driven by individual differences rather than by semantics. Conversely, if we observe low correlations across differently-worded versions of the “same” item pair, then results are more likely to be semantics-influenced. In this study, we use a novel method of paraphrasing scales (described below in Procedure), where we manipulate the semantic relatedness of scale items through the use of large language models (LLMs; Melis et al., 2017; Zhou & Bhat, 2021). This allows for the investigation of how semantic overlap between items affects the observed item response patterns.

Study summary

The current study aims to study the influence of semantic overlap within psychological scales (i.e., the semantic distance between items) on item response patterns (i.e., the empirical correlations between items). We do so by systematically varying the semantic overlap between an existing scale and two ChatGPT-paraphrased versions – one with equal (to original scale) within-scale semantic overlap, and one with lower overlap – for five scales. Then, we predict the *inter-item empirical relationships* using the *inter-item semantic similarities*. Hypotheses listed under Procedure. Data, scripts and other materials are publicly available on: https://osf.io/s2ef4/.

Methods

Throughout the following sections, we will be discussing *two types* of item relationships. We will frequently use the words 'relationships', 'correlations' and 'similarities' in the same sentence, which – although being synonyms of each other – will refer to different concepts in this paper. To help avoid reader confusion, we explicitly disambiguate these terms here. *Semantic similarity*, *semantic overlap* and *semantic relationships* refer to the shared meaning between items and are used interchangeably; the "reverse" term is *semantic distance* (i.e., 'a *decrease* in semantic similarity' means 'an *increase* in semantic distance'). On the other hand, *empirical relationships, inter-item correlations* and (*item*) *response patterns* refer to the empirical correlations between people's responses to two items. Several times, we will present *correlations* between item-pairs' *semantic* and *empirical* relationships, or *correlations* between item-pairs' semantic relationships in one scale version and the same item-pairs' semantic relationships in a different scale version. This kind of easy-to-confuse language use is unavoidable in the context of this study (i.e., correlating correlations). We do our best to keep the language used to discuss the above relationships as clear as possible.

Procedure

To investigate the influence of inter-item semantic overlap on inter-item response patterns, we conducted a within-subjects design study. Participants filled out three versions of a psychological scale – an original (OG), strictly-paraphrased (SP), and loosely-paraphrased (LP) version – over the course of three days. We did not investigate all three versions at the same time for two reasons: a) the number of items across all tests and versions (3 * 48 = 144) was too great for one session, and could have led to low quality data; b) test-takers may have actively recalled their answer to one version of an item and been strongly inclined to respond to a paraphrased version consistently. The presentation of the three versions was counterbalanced.

Before recruiting participants, we paraphrased a scale (see Materials below) using ChatGPT into two new versions – strictly-paraphrased and loosely-paraphrased, using 1) the original scale, and 2) strict/loose-prompts (derived from an exploratory study, pre-registered on: https://osf.io/s2ef4/) as input to the model (see Figures 1 and 2 below). For the strict version, we requested only one paraphrase per item. We then computed sentence embeddings for each item in the original and strict versions using Google's Universal Sentence Encoder (USE; Cer et al., 2018) and, in each version, calculated cosine similarities between item pairs, yielding two symmetric, semantic similarity matrices (SSMs). For the loose paraphrases though, we requested five alternatives per item and selected the combination with the lowest median semantic overlap. This involved computing an SSM for all possible item-alternative combinations, taking the median, and choosing the lowest-internal-similarity combination as the loose version (see Figure 3 for a visual explanation). These two item generation methods can be thought of as two paraphrase-search algorithms –one with low and one with high temperature, where we also optimize for item dissimilarity in the high-temperature algorithm.

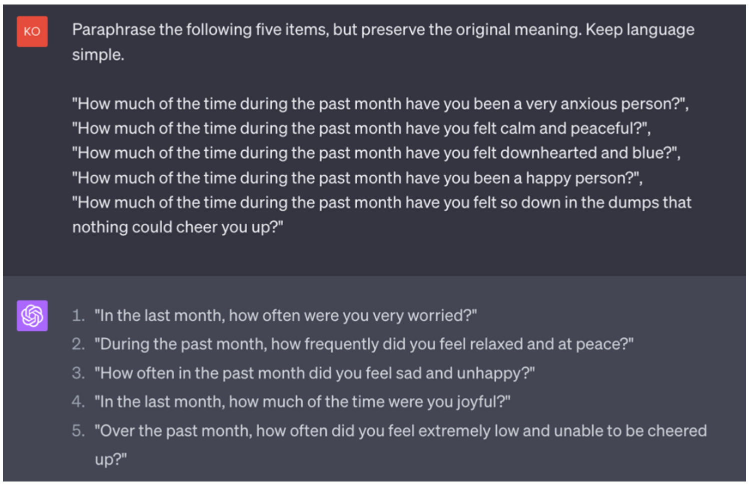
This procedure was conducted for five psychological scales (see Materials below).

After recruitment, participants were informed about the study's procedure and timeline, and provided informed consent. In three sessions over three days, participants filled out each version of the five scales. Each session, they filled out one of the three versions for all five scales. The sessions included two attention checks, totaling 50 items per session. They filled out the three versions in random order during the three measurement occasions.

After the final session, participants were debriefed about the goals of the study.

Figure 1

Strict Scale Paraphrase Using ChatGPT



Paraphrase the following five items, but preserve the original meaning. Keep language simple.

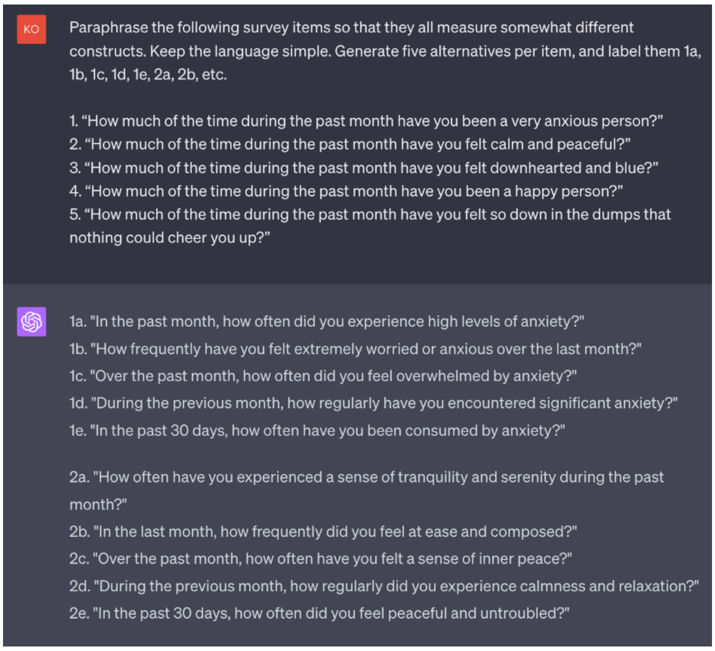
"How much of the time during the past month have you been a very anxious person?",
"How much of the time during the past month have you felt calm and peaceful?",
"How much of the time during the past month have you felt downhearted and blue?",
"How much of the time during the past month have you been a happy person?",
"How much of the time during the past month have you felt so down in the dumps that nothing could cheer you up?"

1. "In the last month, how often were you very worried?"
2. "During the past month, how frequently did you feel relaxed and at peace?"
3. "How often in the past month did you feel sad and unhappy?"
4. "In the last month, how much of the time were you joyful?"
5. "Over the past month, how often did you feel extremely low and unable to be cheered up?"

Note. Model: GPT-4, ChatGPT March 23rd Edition.

Figure 2

Loose Scale Paraphrase Using ChatGPT



Paraphrase the following survey items so that they all measure somewhat different constructs. Keep the language simple. Generate five alternatives per item, and label them 1a, 1b, 1c, 1d, 1e, 2a, 2b, etc.

1. "How much of the time during the past month have you been a very anxious person?"
2. "How much of the time during the past month have you felt calm and peaceful?"
3. "How much of the time during the past month have you felt downhearted and blue?"
4. "How much of the time during the past month have you been a happy person?"
5. "How much of the time during the past month have you felt so down in the dumps that nothing could cheer you up?"

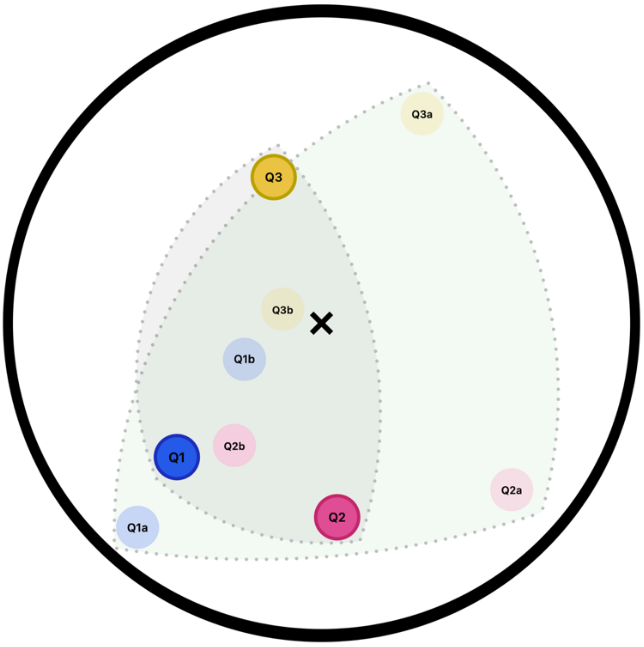
1a. "In the past month, how often did you experience high levels of anxiety?"
1b. "How frequently have you felt extremely worried or anxious over the last month?"
1c. "Over the past month, how often did you feel overwhelmed by anxiety?"
1d. "During the previous month, how regularly have you encountered significant anxiety?"
1e. "In the past 30 days, how often have you been consumed by anxiety?"

2a. "How often have you experienced a sense of tranquility and serenity during the past month?"
2b. "In the last month, how frequently did you feel at ease and composed?"
2c. "Over the past month, how often have you felt a sense of inner peace?"
2d. "During the previous month, how regularly did you experience calmness and relaxation?"
2e. "In the past 30 days, how often did you feel peaceful and untroubled?"

Note. Model: GPT-4, ChatGPT March 23rd Edition.

Figure 3

Visual Aid in Understanding Key Concepts



Note. Consider this figure a 2D representation of "semantic space". Any item (i.e., colored disk) in the circle around '×' is considered as "meaning something about extraversion", while everything outside means something else. Take the extraversion scale Q, with items Q1, Q2 and Q3, and with the distances between the items representing their semantic relatedness. Closer items (Q1 and Q2) overlap more in meaning than distant items (Q1 and Q3). For the loose-paraphrase of scale Q, we generate two alternative per item (Q1a, Q1b, Q2a…). If we want to maximize the distance between item meanings, we would choose items Q1a, Q2a and Q3a. Presumably, responses to the 'a' items are less influenced by shared semantic processing than responses to the original items, since they cover a wider semantic space (green vs. gray area). Moreover, this shows the necessity to consider several item alternatives and optimize for dissimilarity – had we gone with the 'b' items, we would have ended up with a "loose"-paraphrase scale, where the items are *more* semantically related than the original items.

Hypotheses

H1. Item pairs' semantic similarities will be positively related to their inter-item empirical correlations in all three scale versions – OG, SP and LP.

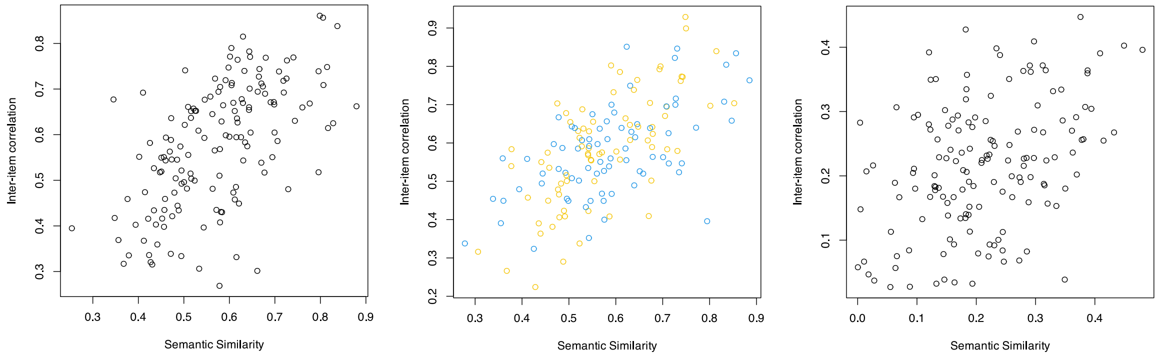
H2. The OG and SP semantic similarities will be equally good predictors of the OG scale and SP scale inter-item empirical correlations, respectively.

H3. The LP similarities are expected to be poorer predictors of the LP inter-item empirical correlations, when compared to the OG semantic similarities and inter-item empirical correlations.

These hypotheses are visualized in Figure 4.

Figure 4

Main Hypotheses Visualized



Note. The three plots depict the potential spread of the data (i.e., simulated data) given the three hypotheses, with the first plot depicting the data under H1, the second – under H2, the third – under H3. Each point in the three visualizations refers to an item *pair*, with the two types of inter-item relationships – semantic and empirical – on the *x* and *y* axes, respectively. H1 predicts a positive association between the two regardless of scale version. H2 predicts that semantic similarity will be as good of a predictor of empirical correlations for the original item-pairs as for the strict item-pairs (i.e. equality). H3 predicts that for the loose item-pairs, semantic similarity will be a poorer predictor of the empirical correlations.

Materials

We chose five measures that covered the following constructs – fear of negative evaluation, fear of missing out, conscientiousness, mind-wandering, and positive expressivity. For each scale, we list a) how many items it has, b) the response scale (all scales were Likert scales; Likert, 1932), and c) the research field the scale originates from.

A. The Positive Expressivity Scale (PE; Barchard, 2001).
   a. Nine items;
   b. 5-point scale;
   c. emotional expressivity.

B. The Fear of Negative Evaluation Scale (FNE; Leary, 1983).
   a. Eleven items;
   b. 5-point scale;
   c. social anxiety.

C. The Fear of Missing Out Scale (FOMO; Abel et al., 2016).
   a. Ten items;
   b. 8-point scale;
   c. digital well-being.

D. The Mind-Wandering: Deliberate & Spontaneous Scales (MW: D&S; Carriere et al., 2013).
   a. Eight items;
   b. 7-point scale;
   c. attention.

E. The Revised NEO Personality Inventory: Conscientiousness, 10-item IPIP revision (NEO-PI: C, IPIP-10; Costa & McCrae, 2008; Goldberg et al., 2006).
   a. Ten items;
   b. 5-point scale;
   c. personality.

Sample

The sample comprised 121 psychology students at the University of Amsterdam. Of these, 11 failed the English fluency check and couldn't proceed to the rest of the survey, and 6 provided no responses. Seven people responded consistently to all items in *at least* one scale (e.g., responding with '2' to all items). Given that the scales had contraindicative items, this type of responding is unlikely to reflect congruent sets of beliefs, hence their data was excluded. This leaves us with 96 participants.

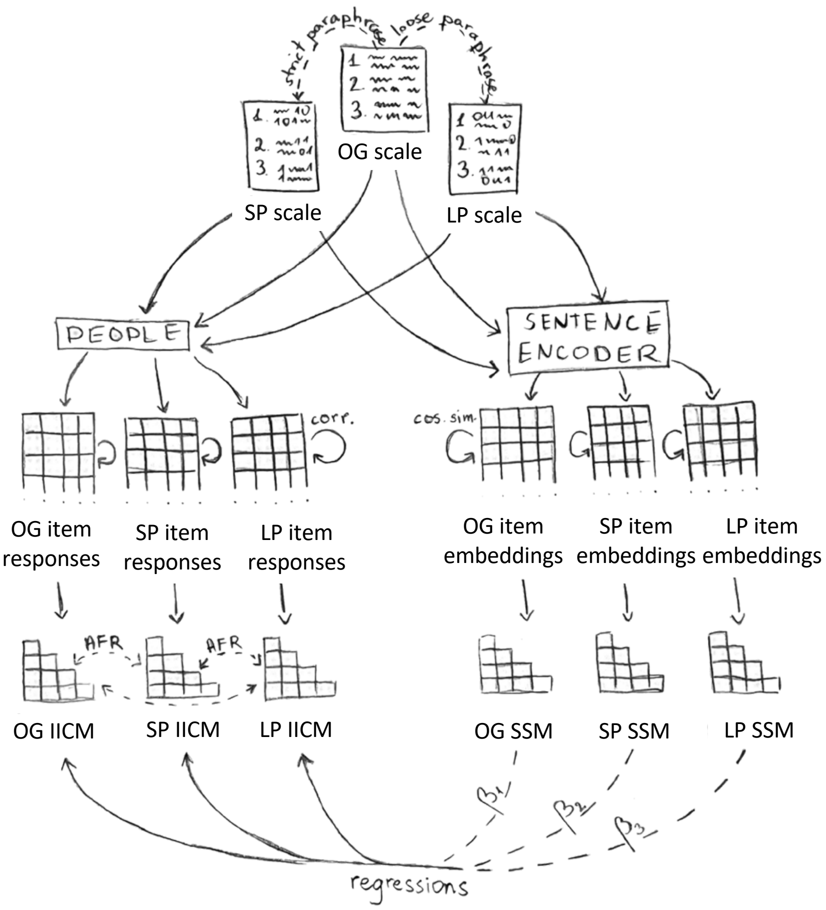
The study had a high rate of drop-out – of the 96 people, only 46 completed all three sets of scale versions, 10 completed two, and 40 completed only one set. Since the presentation of scale version sets was counterbalanced, the original scales were filled out 65 times, the strict-paraphrase versions – 70 times, and the loose versions – 63 times.

Data analysis

Semantic similarities (SSs) were quantified by calculating the cosine similarity between two item embeddings, acquired by passing the items through the USE (Model: USE v4). Relatedly, we calculated inter-item correlations (IICs) by correlating people's responses between items in each scale version (5 scales * 3 versions = 15 matrices). For each scale version, we took the lower triangle of the respective IIC and SS matrices (further referred to as IICM and SSM) , and used these values in our regression analyses. See Figure 5 below for a visualization of the data pipeline.

Figure 5

Schematic of Data Pipeline



Note. Because of the many matrices to keep track of, we drew the data pipeline to improve understanding. For an original (OG) scale, two alternative forms (SP & LP) are created by paraphrasing. For the original and two new versions, people's empirical responses are collected and inter-item correlation matrices (IICMs) are then calculated (left side of drawing). Similarly, item embeddings are computed by passing them through a sentence encoder, and semantic similarity matrices (SSMs) are constructed by calculating the cosine similarity between each item embedding *pair* (right side of drawing). The IICMs and SSMs have the same shape, with each cell describing the same item pair (i.e., the *i*-th, *j*-th IICM cell refers to the same *item pair* as the *i*-th, *j*-th SSM cell). Each of the three IICMs is then regressed on the respective SSM to calculate the SSMs' regression coefficients (i.e., one regression per scale version): a) $OG_{IICM} \sim OG_{SSM}$; b) $SP_{IICM} \sim SP_{SSM}$; c) $LP_{IICM} \sim LP_{SSM}$. These coefficients are then compared to answer H2 (i.e., the OG and SP coefficients will be equal) – and H3 (i.e., the SP coefficient will be smaller than the OG one).

Results

We first conducted a blinded analysis on a shuffled subset of the data (numerical values within columns were shuffled), and then re-ran the analysis script after data collection was complete. The analyses following '*Regressions (H2 & H3)*' are all exploratory.

Descriptives

Five original scales were paraphrased (see Procedure), yielding two new versions – strict and loose – per scale. The scales' mean inter-item semantic similarities are shown in Table 1, while the similarities themselves are presented in a heatmap in Figure 6. After collecting people's responses to the 15 (5 scales * 3 versions) scales, we computed inter-item correlations (IICs; also referred to as item-pairs' *empirical relationships*). The means of these relationships are demonstrated in Table 2.

Table 1

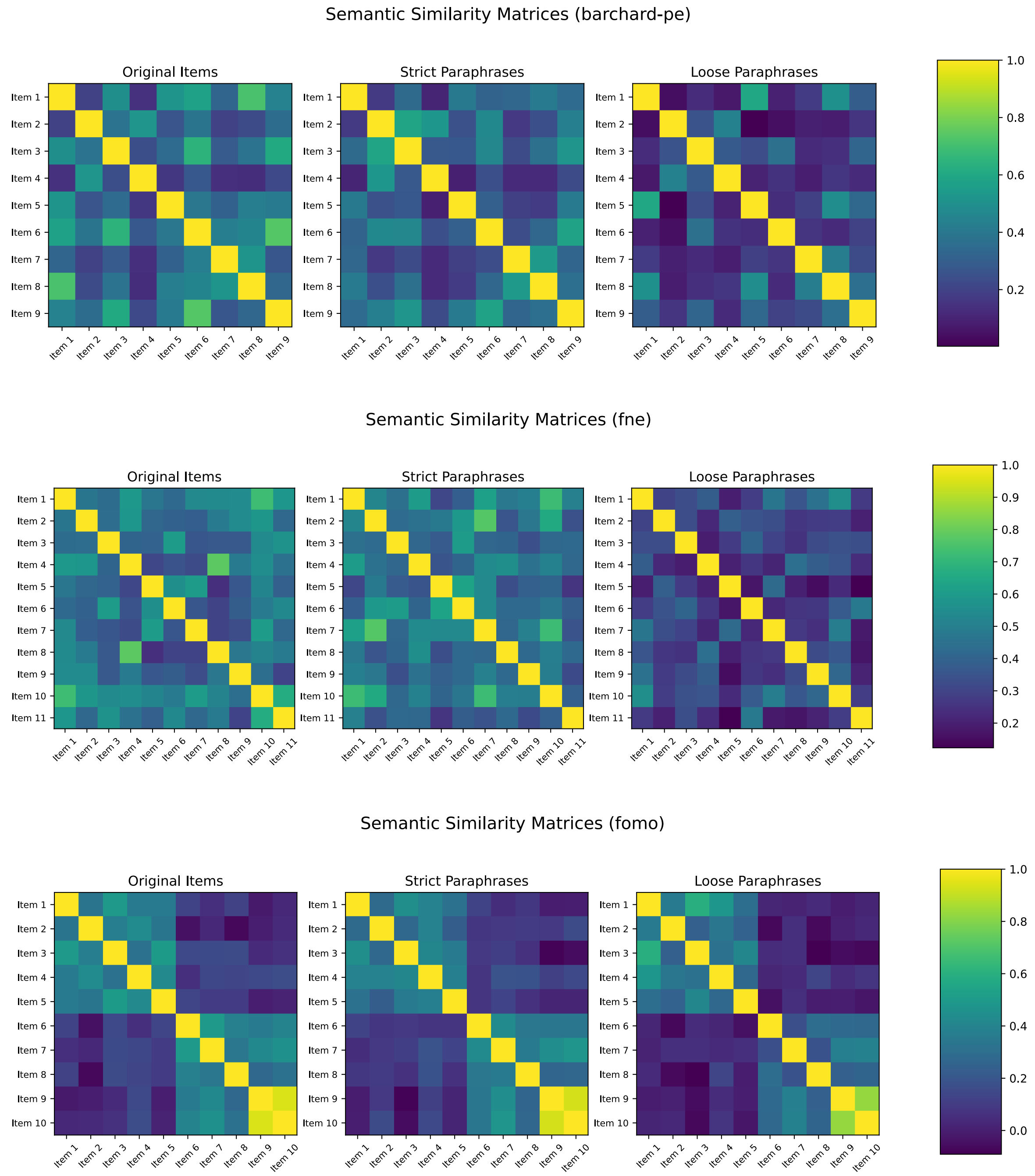
Mean Inter-Item Semantic Similarities per Scale, per Version

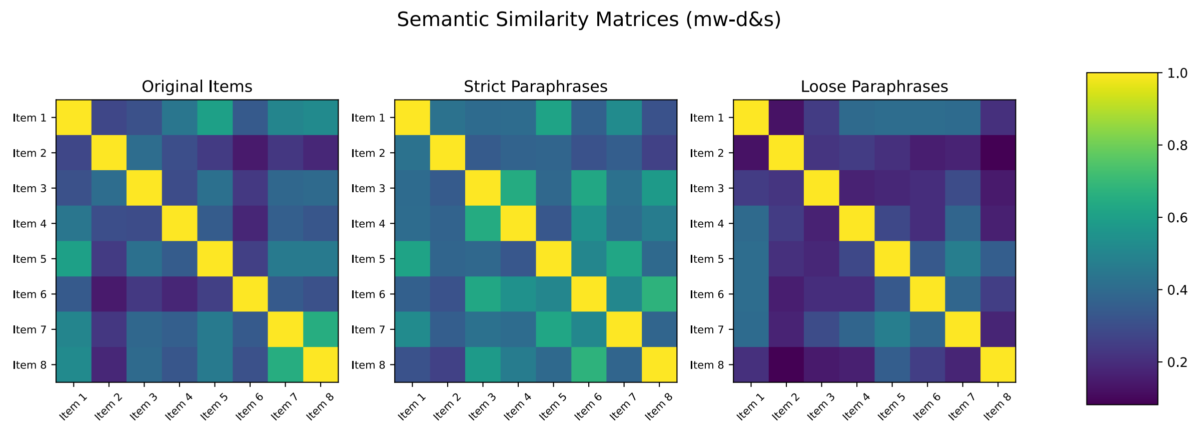
Scale	*mean ss_{OG}*	*mean ss_{SP}*	*mean ss_{LP}*
Overall	**.35 (.16)**	**.36 (.16)**	**.24 (.15)**
PE	.38 (.16)	.32 (.14)	.22 (.15)
FNE	.47 (.11)	.47 (.11)	.30 (.10)
FOMO	.23 (.21)	.22 (.19)	.15 (.19)
MW: D&S	.36 (.12)	.45 (.12)	.26 (.11)
NEO-PI-C	.29 (.09)	.30 (.10)	.22 (.13)

Note. OG = Original, SP = Strict Paraphrase, LP = Loose-Paraphrase. SDs in parentheses.

Figure 6

The Five Scales' Semantic Similarity Matrices





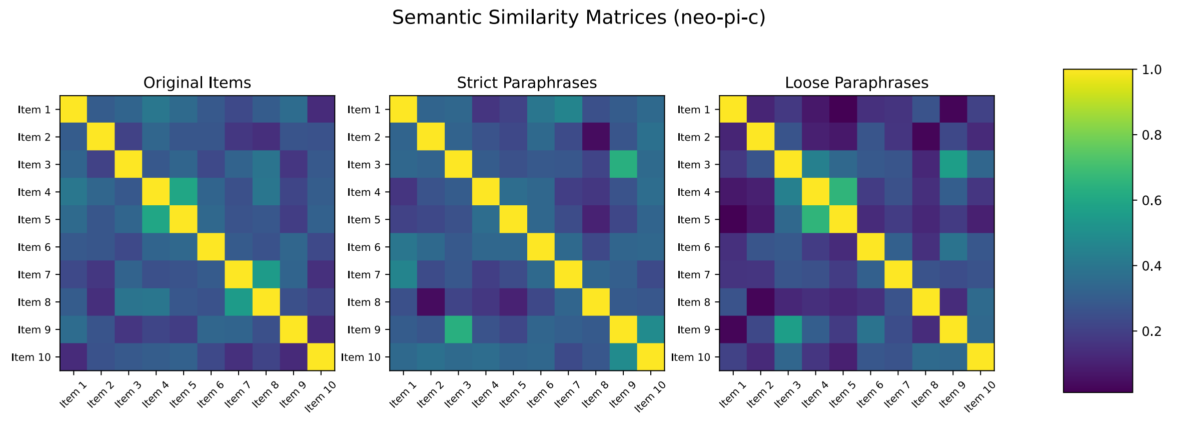


Table 2

Mean Inter-Item Empirical Relationships per Scale, per Version

Scale	*mean* r_{OG}	*mean* r_{SP}	*mean* r_{LP}
Overall	**.39 (.19)**	**.42 (.17)**	**.40 (.19)**
PE	.30 (.17)	.30 (.13)	.30 (.16)
FNE	.50 (.18)	.51 (.13)	.49 (.19)
FOMO	.37 (.20)	.41 (.19)	.37 (.20)
MW: D&S	.38 (.20)	.55 (.14)	.46 (.13)
NEO-PI-C	.37 (.17)	.35 (.13)	.34 (.19)

Note. SDs in parentheses.

Overall, the strict versions' item-pairs' semantic similarities (SP_{SS}) do not differ from the original similarities (OG_{SS}) – a Bayesian paired samples t-test between OG_{SS} and SP_{SS} returned a BF_{01} of

11.89, indicating strong evidence for equality. Whereas the loose versions' item-pairs' semantic similarities (*LP_{SS}*) are lower than the original versions' – a Bayesian paired samples t-test between *OG_{SS}* and *LP_{SS}* returned a $BF_{10} > 100$, indicating extreme evidence in favor of *LP_{SS}* being lower. This is best seen in Figure 6, where the loose item-pairs have "colder" relationships.

CONFIRMATORY ANALYSES

Correlations, semantic and empirical (H1)

H1. Item pairs' semantic similarities will be positively related to their inter-item correlations in all three scale versions – OG, SP and LP.

We conducted Bayesian Pearson's correlation analyses on all the investigated scales to test the correlations between inter-item semantic and empirical relationships. The results are displayed in Table 3, where cell values are the posterior correlation coefficients [and 95% CI] in a given scale version. All three scale versions *overall* show positive correlations between inter-item semantic and empirical relationships – in support of H1; however, 8 of the 15 individual scale versions have correlations where the 95% posterior correlation coefficient CI includes 0.

Table 3

Correlations Between Scales' Inter-Item Semantic Similarities and Empirical Relationships

Scale	*n*	Original (OG)	Strictly-paraphrased (SP)	Loosely-paraphrased (LP)
Overall	**200**	**.37 [.24, .48]**	**.42 [.30, .53]**	**.43 [.30, .53]**
PE	36	.51 [.20, .70]	.25 [-.08, .52] [a]	.62 [.34, .78]
FNE	55	.01 [-.25, .27] [a]	.02 [-.24, .26] [a]	.06 [-.21, .31] [a]
FOMO	36	.51 [.21, .70]	.50 [.20, .70]	.51 [.21, .71]
MW: D&S	28	.02 [-.34, .37] [a]	.18 [-.20, .50] [a]	.17 [-.20, .49] [a]

Scale	*n*	Original (OG)	Strictly- paraphrased (SP)	Loosely- paraphrased (LP)
NEO-PI-C	45	.48 [.20, .66]	.21 [-.08, .47]^{a}	.44 [.16, .64]

Note. ‘*n*’ refers to the number of inter-item relationships in a scale. For example, the PE scale has 9 items, meaning there are 36 relationships in total. Posterior means [and 95% CIs] shown. ‘^{a}’ marks 95% CIs that include 0.

Regressions (H2 & H3)

H2. The OG and SP semantic similarities will be equally good predictors of the OG scale and SP scale inter-item correlations, respectively.

H3. The LP similarities are expected to be poorer predictors of the LP inter-item correlations, when compared to the OG semantic similarities and inter-item correlations.

Since we wanted to compare the predictive power of the different scale versions’ semantic similarities, we conducted Bayesian linear regressions on the three types of investigated scales – original, strictly- and loosely-paraphrased (OG, SP, LP). We used weakly informative priors for the β and σ parameters in all regressions, based on a blogpost by Andrew Gelman (https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations):

$$\beta \sim Student\ t\ (3, 0, 2.5)$$

$$\sigma \sim Half\ Cauchy\ (0, 2.5)$$

In the three regression models, the empirical inter-item correlations were regressed on their respective semantic similarities (see Table 4 and Figure 7):

$$\textit{Model 1.} \quad OG_{IICM} \sim OG_{SSM}$$

$$\textit{Model 2.} \quad SP_{IICM} \sim SP_{SSM}$$

$$\textit{Model 3.} \quad LP_{IICM} \sim LP_{SSM}$$

Posterior predictive checks from the fitted models showed that the predicted data followed the observed data well (Figure 10). We then drew 8000 regression coefficient samples each from the respective posterior distributions and calculated difference distributions in a given comparison (e.g., β_{OG} minus β_{SP}). These difference distributions' CIs are displayed in Table 7. The posterior β_{OG} and β_{SP} difference distribution's CI lies at [-.20, .20] – centered at 0 – meaning no difference in predictive power between the OG and SP scale versions' semantic similarities. This is in support of H2. However, the posterior β_{OG} and β_{LP} difference distribution's CI also contains 0, meaning no support for H3.

Table 4

Regressions of the Three Scale Versions' IICs on their Respective SSs.

Model	Effect	Estimate	Est. Error	95% CI		$\check{R}$
				LB	*UB*	
$\boldsymbol{OG_{IICM} \sim OG_{SSM}}$	Intercept	.24	.03	.18	.30	1.00
	Semantic similarities	**.44**	.08	**.29**	**.59**	1.00
	Sigma	.18	.01	.17	.20	1.00
$\boldsymbol{SP_{IICM} \sim SP_{SSM}}$	Intercept	.27	.03	.22	.32	1.00
	Semantic similarities	**.44**	.07	**.31**	**.58**	1.00
	Sigma	.16	.01	.14	.17	1.00
$\boldsymbol{LP_{IICM} \sim LP_{SSM}}$	Intercept	.26	.02	.22	.31	1.00
	Semantic similarities	**.56**	.09	**.39**	**.73**	1.00
	Sigma	.17	.01	.16	.19	1.00

Figure 7

Scale Versions' Regression Plots

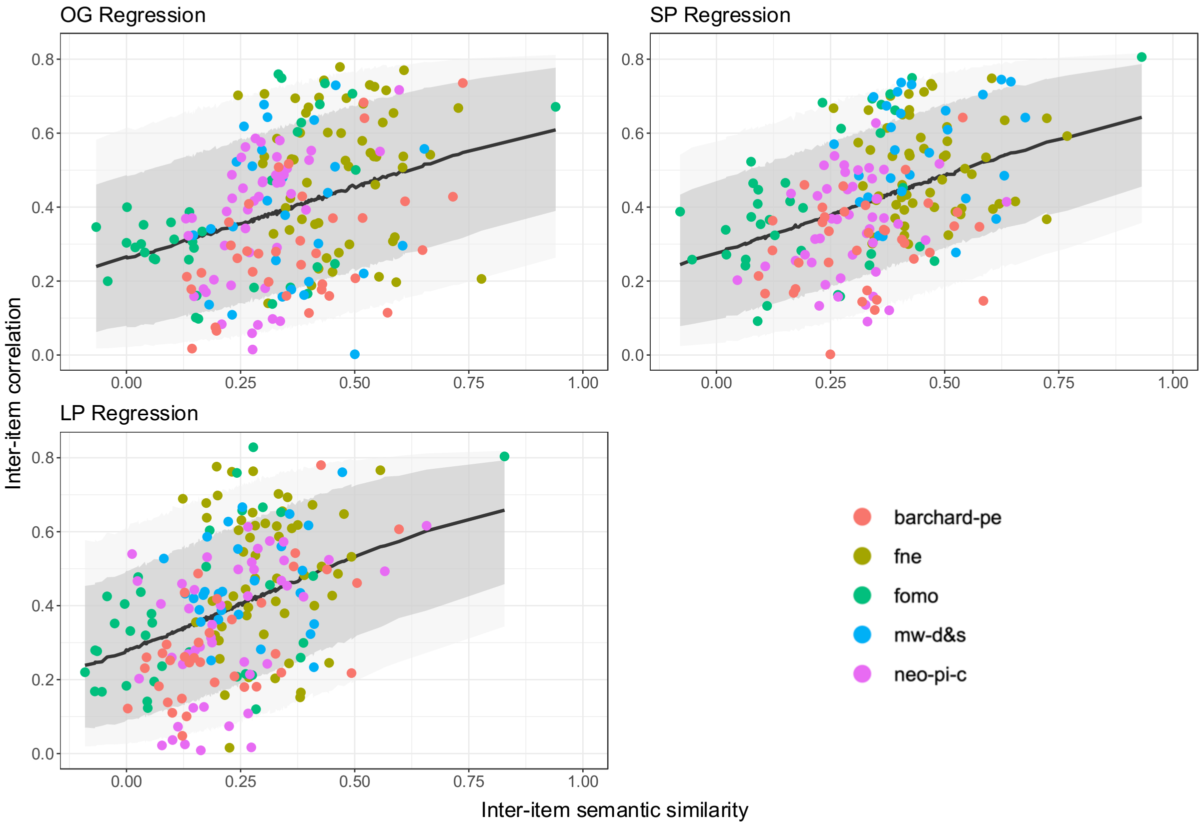
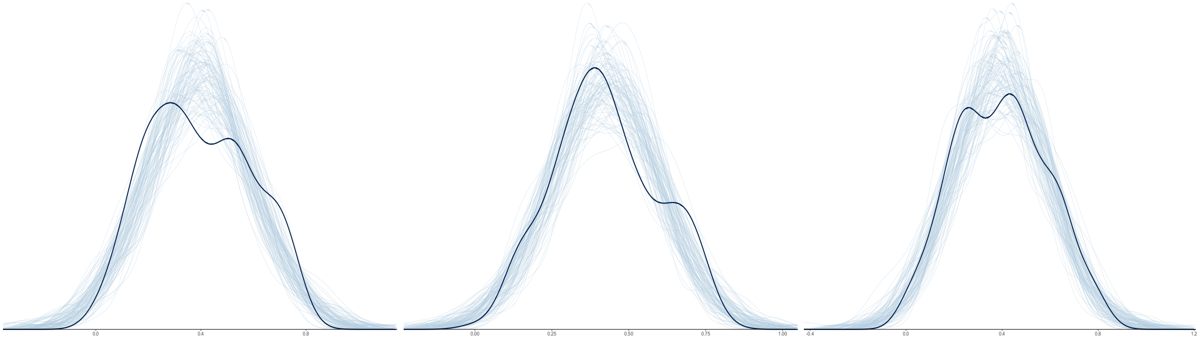


Figure 8

Posterior Predictive Checks (Models 1-3)



Note. Dark blue lines show the observed data from the three models. Light blue lines indicate model-predicted data.

Table 5

Posterior Difference Distributions' 95% CIs

	β_{OG}	β_{SP}	β_{LP}
β_{OG}	–		
β_{SP}	[-.20, .20] ^{a}	–	
β_{LP}	[-.34, .11] ^{a}	[-.34, .09] ^{a}	–

Note. Read column, then row – the posterior β_{OG} and β_{SP} difference distribution's 95% CI lies at [-.20, .20]. '^{a}' marks 95% CIs that include 0.

Comparing regression coefficients from models with different independent and dependent variables in a manner like above would generally be considered dubious at best; however, in our study both the independent and dependent variables are each of the same units, and have near-equal variance. Standardizing the variables – so that all have means of zero and variances of one – yields no numerical difference in the regression coefficients' and/or differences distributions' CIs.

EXPLORATORY ANALYSES

The following analyses are all post-hoc and do not bring the validity of our confirmatory analyses into question; however, they provide us with important nuance for our Discussion section, hence their inclusion in the main body.

Correlations, between-version

To see how well the original inter-item semantic similarities were preserved across the paraphrased versions following the paraphrasing procedure, we calculated correlations between item-pair *semantic similarities* in the different scale versions (e.g., the correlation between the OG PE and SP PE scales' inter-item semantic similarities). This can be thought of as finding the between-version correlation of the matrices in Figure 6. As seen in Table 6, most scales display medium-high between-version relationships, except for the NEO-PI-C scale, where the OG version's semantic relationships are near-unrelated to the two other versions'.

Table 6

Between-Version Semantic Similarity Correlations

Scale	OG-SP	OG-LP	SP-LP
Overall	**.70**	**.66**	**.67**
PE	.77	.57	.48
FNE	.48	.57	.57
FOMO	.94	.94	.95
MW: D&S	.09	.43	.25
NEO-PI-C	-.07	.07	.50

Internal consistencies

To study consistency between people's responses to all the studied scales, we calculated internal consistency scores (Cronbach's Alpha) – shown in Table 7. All scales exhibit at least "acceptable" (i.e., ≥ .7) internal consistency (with the exception of the OG PE scale, where a portion of the density dips below .7), with most scales hovering around "good" (i.e., ≥ .8) scores.

Table 7

Internal Consistencies (Cronbach's Alpha) of the Investigated Scales

Scale	Original (OG)	Strictly-paraphrased (SP)	Loosely-paraphrased (LP)
PE	.77 [.698, .85]	.78 [.70, .85]	.78 [.71, .86]
FNE	.92 [.89, .94]	.92 [.89, .94]	.88 [.84, .92]

Scale	Original (OG)	Strictly-paraphrased (SP)	Loosely-paraphrased (LP)
FOMO	.83 [.78, .90]	.85 [.798, .90]	.84 [.78, .90]
MW: D&S	.82 [.75, .88]	.90 [.87, .94]	.87 [.82, .91]
NEO-PI-C	.85 [.80, .90]	.81 [.75, .87]	.83 [.77, .89]

Note. Posterior means [and 95% CIs]. Rounded to two decimals unless the third decimal matters.

Alternative-form reliabilities (AFRs)

To see how well the paraphrased scales function as alternative forms of the original scales, we calculated AFRs. We did so by calculating participants' sum scores on all the scales they filled out, and then correlating these scores across the different scale versions (e.g., for a participant who filled out the OG and SP scales, we calculated all their sum scores, and then correlated all the OG scales' scores with their analogous SP scores). The results are displayed in Table 8 and Figure 9.

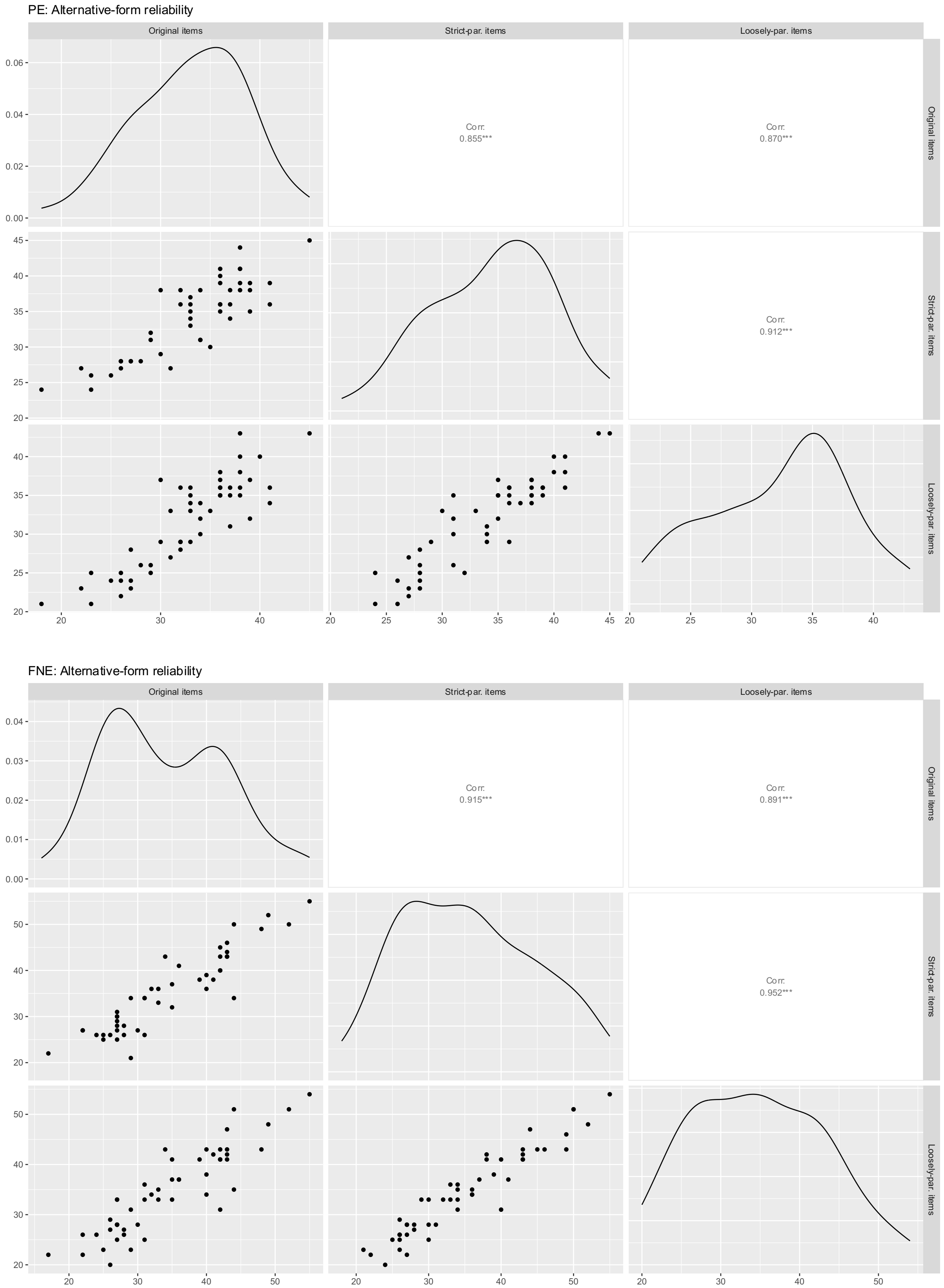
Table 8

Mean AFRs per Scale

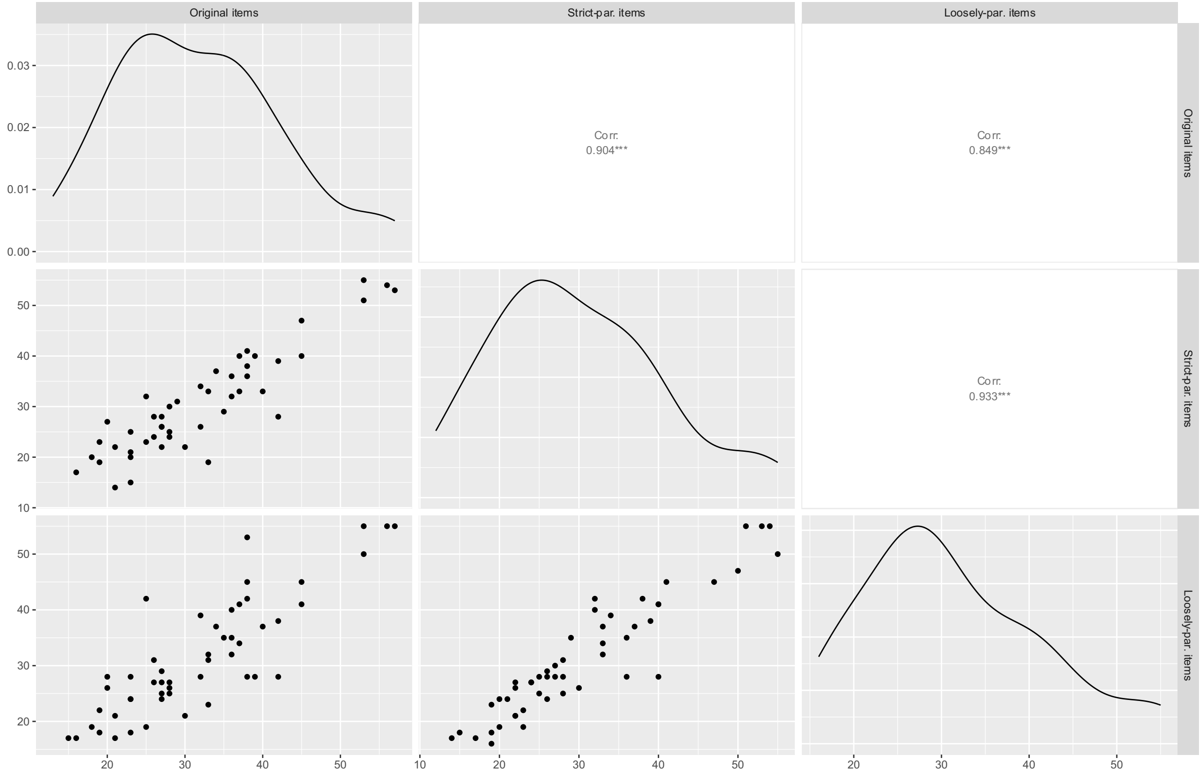
Scale	*mean AFR*
PE	.879
FNE	.919
FOMO	.895
MW: D&S	.893
NEO-PI-C	.908

Figure 9

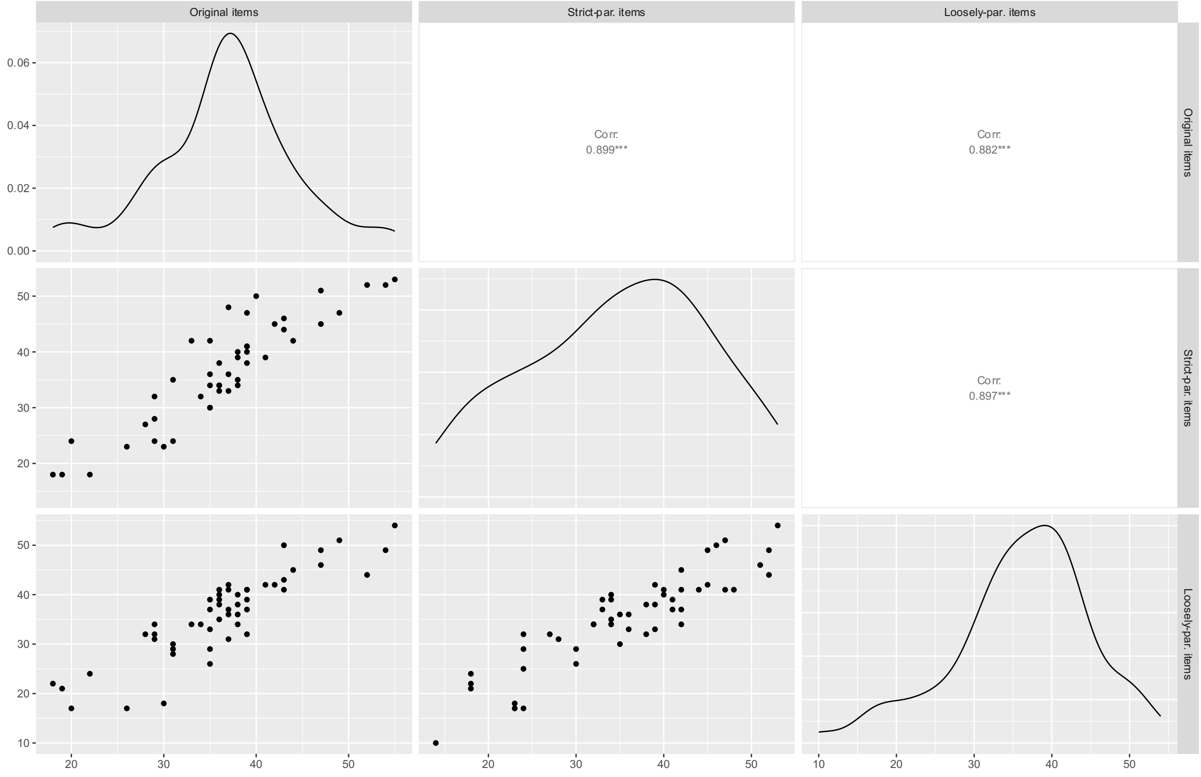
AFRs of the Five Scales



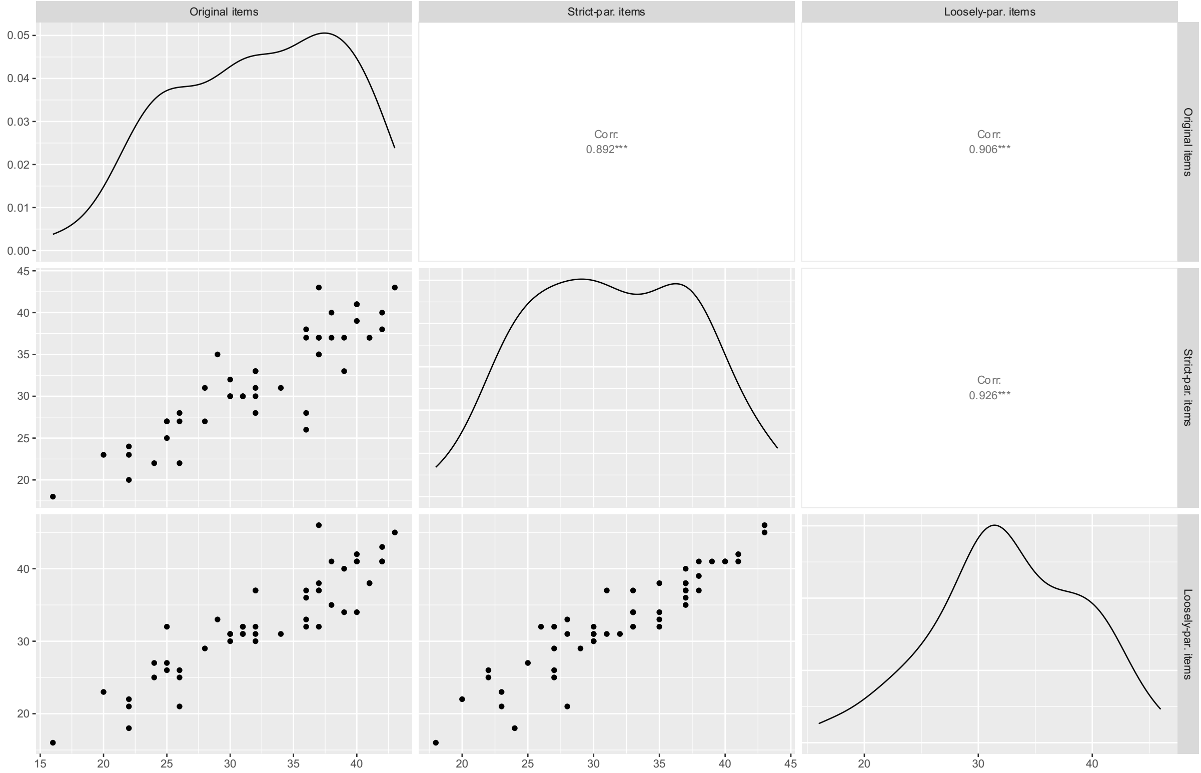
FOMO: Alternative-form reliability



MW: D&S: Alternative-form reliability



C: Alternative-form reliability



Discussion

Our goal in this study was to better understand the influence that semantic overlap between items in psychological scales has on people's item response patterns. We systematically manipulated the amount of within-scale semantic overlap by paraphrasing five existing, commonly-used scales using GPT-4, according to a strict-paraphrase method and a loose-paraphrase method (see Appendix A for the investigated scales with items). With our strict-paraphrase method, we paraphrased the scales so that they were worded differently but that the amount of within-scale overlap was the same as with the original scale. Whereas with our loose method, we successfully lowered the amount of within-scale semantic overlap; in other words, the loose items were "pulled apart" from each other in terms of meaning. After participants filled in the original, strict and loose scale versions – we predicted the *empirical correlations* between all item-pairs using the same item-pairs' *semantic similarities*, across the three scale versions (see Figure 5 for visual schematic). We found a positive association between item pairs' semantic and empirical relationships in all three types of scale version – original, strict and loose (as predicted by H1). Moreover, we found that the *original* and *strict* inter-item semantic similarities were equally good predictors of the *original* and *strict* inter-item empirical correlations, respectively (as predicted by H2). This is in line with our theory – despite changes in item wording, the overall semantic similarity between items in the original and strict scales was the same, and so we did not expect any change (i.e., a strengthening/weakening) in the overall item response patterns. With H3, we predicted that the *loose* semantic similarities would be *poorer* predictors of the *loose* inter-item empirical correlations. This is because we thought people would respond less consistently to the "pulled apart" items due to less overlap in meaning. However, the loose similarities predicted people's loose response patterns *as well* as the original or strict similarities did with the original and strict scales. Hence, we found no evidence that *increasing* semantic distance between scale items results in *weaker* inter-item correlations. To further back this up, we calculated internal consistencies for all investigated scales and found that people's responses were *as consistent* in all scale versions (see Table 7). Again, we would have expected the loose scales to have *lower* internal consistency. Finally, we found that the paraphrased scales were excellent alternate forms of the original scales (regardless if strict or loose). Altogether, these findings suggest that semantic overlap between items plays a weaker role in "inflating" empirical inter-item correlations than we initially thought. In terms of real-life practice, we now believe that regardless of whether a researcher used one of our original, strict or loose scale versions to investigate a construct of interest, they would likely observe equivalent results.

Limitations

Our results rest on the assumption that our method of increasing the semantic distance between items truly worked. We increased this distance by selecting item alternatives that overall (i.e., in the entire scale) had the lowest amount of shared meaning (see Procedure for a detailed rundown). However, it is possible that our observed manipulation of semantic distance is an artefact of the item embedding/comparison procedure, and does *not* reflect a true decrease in shared meaning. Syntactic changes such as swapping the places of two words or clauses in a sentence (in a way that does not alter meaning) will result in decreased semantic similarity when compared to the original sentence. The same goes for swapping words for equivalent synonyms. While these changes are irrelevant regarding a sentence's meaning, they still influence said sentence's embeddings, and hence, the distance with other sentences. Hence, we believe our loosely-paraphrased scales are contaminated by these "superficial" changes in semantic distance that do not reflect any real difference in shared meaning; however, it is difficult to estimate to what extent our manipulation was "superficial" as opposed to "deep" (i.e., reflecting *real* shifts in item meaning). Quantifying and studying the (co-)influences of the two, or developing a method that measures only "deep" semantic shifts is a logical next step for the current line of research

Another assumption that the above interpretations depend on is that participants responses to the different scale versions were independent. Meaning, that memory of their response to OG PE Item #1 had no influence on their response to SP PE Item #1 the next day. If participants remembered their responses well enough across the testing days, then *memory* could be driving consistency across versions up. Therefore, we re-ran our analyses using only the first block of scales that people filled out (see Appendix B). The evidence in favor / against our hypotheses did not change, hence we rule out memory as a confound.

Scale development

Psychology students are taught several item/scale construction guidelines during their first psychometrics courses under the notion that breaking said guidelines results in poorer scales. In our study, we didn't prompt GPT-4 to follow any guidelines other than "Keep the language simple", and so several of the AI-generated items violated common guidelines – e.g., "avoid negative statements", "state the situational clause before the behavioral clause", "write items to sixth grade reading level", etc. (Comrey, 1988; Kline, 2015). However, these violations didn't result in "poorer" scale properties in terms of their AFRs and internal consistencies. Meaning that using the paraphrased scales over the original ones for their intended purposes would likely *not* result in different interpretations regarding people's/groups' responses. These results spell good news for scale developers, since we provide evidence that some item wording conventions don't matter as much as commonly thought. Including esoterically worded items still produces empirical inter-item relationships that are as strong and "unaffected" by odd phrasing (see Appendix A and Tables 2 and 7). This makes us question to what extent these common guidelines are truly helpful, and where on the spectrum from empirically-supported to folk wisdoms to "alternative medicines" they lie. Our skepticism is mild though, primarily due to the low sample size in the current study. Therefore, future research could address this skepticism by comparing "rule-breaking" scales to original ones in a study similar to ours.

Conclusion

The present study sought to better understand the role that semantic overlap between items in psychological scales has on item response patterns. We systematically varied the amount of within-scale semantic overlap in five commonly-used scale by paraphrasing them using GPT-4. We found that increasing semantic distance between items did *not* result in weaker inter-item empirical relationships, easing our suspicions that high semantic consistency between items was inflating the empirical inter-item correlations found in psychological literature. However, we suspect that our method of manipulating semantic overlap was contaminated by "superficial" embedding artefacts that do not have an influence on the actual shared meaning between items. Based on these findings, future research should delve deeper into separating "superficial" and "deep" changes in scale item semantics, as well as further questioning the usefulness of commonly used item construction guidelines.

References

Abel, J. P., Buff, C. L., & Burr, S. A. (2016). Social Media and the Fear of Missing Out: Scale Development and Assessment. *Journal of Business & Economics Research (JBER)*, *14*(1), 33–44. https://doi.org/10.19030/jber.v14i1.9554

Arnulf, J. K., & Larsen, K. R. (2021). Semantic and ontological structures of psychological attributes. In *Measuring and Modeling Persons and Situations* (pp. 69–101). Elsevier. https://doi.org/10.1016/B978-0-12-819200-9.00013-2

Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., & Bong, C. H. (2014). Predicting Survey Responses: How and Why Semantics Shape Survey Statistics on Organizational Behaviour. *PLOS ONE*, *9*(9), e106361. https://doi.org/10.1371/journal.pone.0106361

Barchard, K. A. (2001). *Emotional and social intelligence: Examining its place in the nomological network*. https://doi.org/10.14288/1.0090848

Bowers, K. S. (1973). Situationism in psychology: An analysis and a critique. *Psychological Review*, *80*(5), 307–336.

Carriere, J. S. A., Seli, P., & Smilek, D. (2013). Wandering in both mind and body: Individual differences in mind wandering and inattention predict fidgeting. *Canadian Journal of Experimental Psychology / Revue Canadienne de Psychologie Expérimentale*, *67*(1), 19–31. https://doi.org/10.1037/a0031438

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., & Kurzweil, R. (2018). *Universal Sentence Encoder* (arXiv:1803.11175). arXiv. http://arxiv.org/abs/1803.11175

Chowdhary, K. R. (2020). Natural Language Processing. In K. R. Chowdhary (Ed.), *Fundamentals of Artificial Intelligence* (pp. 603–649). Springer India. https://doi.org/10.1007/978-81-322-3972-7_19

Comrey, A. L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology*, *56*(5), 754–761. https://doi.org/10.1037/0022-006X.56.5.754

Cook, D. A., & Beckman, T. J. (2006). Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application. *The American Journal of Medicine*, *119*(2), 166.e7-166.e16. https://doi.org/10.1016/j.amjmed.2005.10.036

Coombs, C. H. (1964). *A theory of data.* (pp. xviii, 585). Wiley.

Costa, P. T., & McCrae, R. R. (2008). The Revised NEO Personality Inventory (NEO-PI-R). In *The SAGE Handbook of Personality Theory and Assessment: Volume 2—Personality Measurement and Testing* (pp. 179–198). SAGE Publications Ltd. https://doi.org/10.4135/9781849200479.n9

Drost, E. A. (2011). Validity and reliability in social science research. *Education Research and Perspectives*, *38*(1), 105–123. https://doi.org/10.3316/informit.491551710186460

Gliem, J. A., & Gliem, R. R. (2003). *Calculating, Interpreting, And Reporting Cronbach's Alpha Reliability Coefficient For Likert-Type Scales*. https://scholarworks.iupui.edu/handle/1805/344

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*(1), 84–96. https://doi.org/10.1016/j.jrp.2005.08.007

Hofer, G. (2020). *Sample size and stability of correlation coefficients: A replication of Schönbrodt &amp; Perugini (2013)* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/ygm57

Joos, M. (1950). Description of Language Design. *The Journal of the Acoustical Society of America*, *22*(6), 701–707. https://doi.org/10.1121/1.1906674

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall.

Kline, P. (2015). *A Handbook of Test Construction (Psychology Revivals)* (0 ed.). Routledge. https://doi.org/10.4324/9781315695990

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

Leary, M. R. (1983). A Brief Version of the Fear of Negative Evaluation Scale. *Personality and Social Psychology Bulletin*, *9*(3), 371–375. https://doi.org/10.1177/0146167283093007

Maul, A. (2017). Rethinking Traditional Methods of Survey Validation. *Measurement: Interdisciplinary Research and Perspectives*, *15*(2), 51–69. https://doi.org/10.1080/15366367.2017.1348108

McNemar, Q. (1955). Psychological statistics. *Journal of Consulting Psychology*, *19*(2), 155–155. https://doi.org/10.1037/h0039325

Melis, G., Dyer, C., & Blunsom, P. (2017). *On the State of the Art of Evaluation in Neural Language Models* (arXiv:1707.05589). arXiv. http://arxiv.org/abs/1707.05589

Meyer, R. D., Dalal, R. S., & Hermida, R. (2010). A Review and Synthesis of Situational Strength in the Organizational Sciences. *Journal of Management*, *36*(1), 121–140. https://doi.org/10.1177/0149206309349309

Michell, J. (1994). Measuring Dimensions of Belief by Unidimensional Unfolding. *Journal of Mathematical Psychology*, *38*(2), 244–273. https://doi.org/10.1006/jmps.1994.1016

Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review*, *80*(4), 252–283. https://doi.org/10.1037/h0035002

Murphy, K. R. (Ed.). (1996). *Individual differences and behavior in organizations* (1st ed). Jossey-Bass Publishers.

Nimon, K., Shuck, B., & Zigarmi, D. (2016). Construct Overlap Between Employee Engagement and Job Satisfaction: A Function of Semantic Equivalence? *Journal of Happiness Studies*, *17*(3), 1149–1171. https://doi.org/10.1007/s10902-015-9636-6

Nunnally, J. C. (1978). An Overview of Psychological Measurement. In B. B. Wolman (Ed.), *Clinical Diagnosis of Mental Disorders: A Handbook* (pp. 97–146). Springer US. https://doi.org/10.1007/978-1-4684-2490-4_4

Sadler, P., & Woody, E. (2003). Is who you are who you're talking to? Interpersonal style and complementarily in mixed-sex interactions. *Journal of Personality and Social Psychology*, *84*(1), 80–96. https://doi.org/10.1037/0022-3514.84.1.80

Şahin, M. D. (2021). Effect of Item Order on Certain Psychometric Properties: A Demonstration on a Cyberloafing Scale. *Frontiers in Psychology*, *12*, 590545. https://doi.org/10.3389/fpsyg.2021.590545

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. http://arxiv.org/abs/1706.03762

Wood, D., Read, S. J., Harms, P. D., & Slaughter, A. (Eds.). (2021). *Measuring and modeling persons and situations*. Academic Press.

Zhou, J., & Bhat, S. (2021). Paraphrase Generation: A Survey of the State of the Art. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 5075–5086. https://doi.org/10.18653/v1/2021.emnlp-main.414

Appendix A: The Five Scales and their Paraphrased Versions

Below are the five scales we used in the current study, along with their strictly- and loosely-paraphrased versions.

Table A1

PE Scale

Original	Strict	Loose
I express my affection physically.	I show my love through touch.	I communicate my love through physical contact.
I laugh out loud if something is funny.	I openly laugh when something makes me amused.	If I find something amusing, you'll likely hear me laugh.
I express my happiness in a childlike manner.	I display my joy in a playful way.	My expressions of happiness often resemble the uninhibited joy of a child.
I sometimes laugh out loud while reading, watching videos, or streaming shows.	Sometimes, I can't help but laugh out loud when I read, watch videos, or binge-watch TV shows.	It's not unusual for me to laugh out loud while watching or reading something.
I hug my close friends.	I embrace my best friends.	Physical closeness, like hugging, is one way I connect with my close friends.
I show my feelings when I'm happy.	I express my emotions when I feel joyful.	When I am happy, I make it known.
I find it difficult showing people that I care about them.	It's hard for me to show others that I value them.	It's challenging for me to show others how much I value them.
I have difficulty showing affection.	It's tough for me to demonstrate love.	I find it hard to show my affectionate side.
I keep my happy feelings to myself.	I tend to hold back my happiness.	I keep my positive emotions within myself.

Table A2

FNE Scale

Original	Strict	Loose
I worry about what other people will think of me even when I know it doesn't make any difference.	I often stress about others' opinions of me, even when it doesn't matter.	I often stress over others' opinions of me, regardless of their significance.
I am unconcerned even if I know people are forming an unfavorable impression of me.	I don't care if people have a bad opinion of me.	The idea of people forming a negative image of me doesn't cause me any concern.
I am frequently afraid of other people noticing my shortcomings.	I often fear people will notice my flaws.	The fear of others observing my faults frequently haunts me.
I rarely worry about what kind of impression I am making on someone.	I don't usually worry about the impression I leave on others.	I infrequently fret over the kind of influence I might be having on someone.
I am afraid that others will not approve of me.	I'm scared others won't accept me.	I'm afraid of not receiving approval from people.
I am afraid that people will find fault with me.	I'm scared people will criticize me.	I worry that people might identify and highlight my mistakes.
Other people's opinions of me do not bother me.	I'm not bothered by what others think of me.	I remain indifferent to people's opinions of me.
When I am talking to someone, I worry about what kind of impression I make.	When talking to someone, I worry about the impression I give.	While talking to someone, I stress over the kind of impression I might be leaving.
If I know someone is judging me, it has little effect on me.	Even if someone judges me, it doesn't affect me much.	Even when I know someone is assessing me, it doesn't usually impact me.
Sometimes I think I am too concerned with what other people think of me.	I sometimes feel I care too much about others' opinions.	At times, I think I give too much importance to how others view me.
I often worry that I will say or do the wrong things.	I frequently worry about making mistakes in what I say or do.	I regularly worry about making a mistake in my words or actions.

Table A3

FOMO Scale

Original	**Strict**	**Loose**
I take a positive attitude toward myself.	I have a positive view of myself.	I generally have a positive view of myself.
On the whole, I am satisfied with myself.	Overall, I am content with who I am.	Overall, I am pleased with myself.
I feel I have a number of good qualities.	I believe I possess several good traits.	I think I have quite a few positive attributes.
All in all, I am inclined to feel that I am a failure.	Generally, I tend to think I am unsuccessful.	Generally, I lean towards believing that I am not successful.
I feel I do not have much to be proud of.	I don't think I have much to be proud of.	I believe I lack significant accomplishments to be proud of.
Do you feel uncomfortable meeting new people?	Does meeting new people make you uneasy?	Does encountering new people cause you distress?
How frequently are you troubled by shyness?	How often does shyness bother you?	How frequently do you grapple with timidity?
When in a group of people, do you have trouble thinking of the right things to talk about?	Do you find it difficult to think of topics to discuss in a group?	When surrounded by others, do you find it challenging to come up with conversation topics?
Assume you are unable to check social media when you want to. Generally, how frequently do you feel frightened?	If you can't access social media when desired, how often do you feel scared?	If you are unable to access social media when desired, how often do you feel scared?
Assume you are unable to check social media when you want to. Generally, how frequently do you feel nervous?	If you can't access social media when desired, how often do you feel anxious?	If you can't check social media when you want, how often do you feel anxious?

Table A4

MW: D&S Scale

Original	Strict	Loose
I allow my thoughts to wander on purpose.	I purposely let my mind drift.	I consciously permit my thoughts to roam freely.
I enjoy mind-wandering.	I like letting my mind roam freely.	I find satisfaction in mind-wandering.
I find mind-wandering is a good way to cope with boredom.	I use daydreaming as a way to deal with dullness.	I consider mind-wandering a beneficial strategy to combat boredom.
I allow myself to get absorbed in pleasant fantasy.	I permit myself to become engrossed in enjoyable daydreams.	I grant myself permission to get engulfed in delightful daydreams.
I find my thoughts wandering spontaneously.	My thoughts often drift on their own.	Often, I find my mind wandering without my control.
When I mind-wander my thoughts tend to be pulled from topic to topic.	My thoughts shift from one subject to another when I daydream.	In my mind-wandering moments, my thoughts hop between topics.
It feels like I don’t have control over when my mind wanders.	It seems like I can't control when my thoughts start drifting.	I feel I lack control over the timing of my mind-wandering.
I mind wander even when I’m supposed to be doing something else.	I daydream even when I should be focusing on other tasks.	My mind drifts even in the midst of other obligations.

Table A5

NEO-PI-C Scale

Original	Strict	Loose
I am always prepared.	I'm constantly ready.	Being unprepared is not in my nature.
I pay attention to details.	I focus on the little things.	I notice the small things.
I get chores done right away.	I complete tasks immediately.	I tackle tasks immediately.
I carry out my plans.	I execute my strategies.	I execute my strategies effectively.
I make plans and stick to them.	I create and follow my plans.	I create strategies and adhere to them.
I waste my time.	I squander my time.	I often fritter away my time.
I find it difficult to get down to work.	It's hard for me to start working.	I find it hard to settle into work.
I do just enough work to get by.	I only do the minimum required.	I tend to do only what's necessary to meet expectations.
I don't see things through.	I fail to finish tasks.	I often leave tasks midway.
I shirk my duties.	I avoid my responsibilities.	I'm prone to neglecting my obligations.

Appendix B: First-Block-Only Results

To investigate whether memory could have confounded our results, we reran our confirmatory analyses using only the first block of scale versions that a participant filled out. For example, if a participant got the scale order 'SP, LP, OG' across the three measurement days, then we consider only their SP scales' data. By doing this, we conduct a "post-hoc between-subjects" analysis, where memory cannot play a role in people's responses (since we are investigating only the *first* set of scales they completed).

Correlations, semantic and empirical (H1)

Overall, the same pattern of results holds when retesting H1. When looking at the individual scales, nearly the same pattern of results holds, with the exception of the SP FOMO scale's 95% CI now including 0.

Table B1 (H1)

Correlations Between Scales' Inter-Item Semantic Similarities and Empirical Relationships

Scale	*n*	Original (OG)	Strictly-paraphrased (SP)	Loosely-paraphrased (LP)
Overall	**200**	**.32 [.19, .44]**	**.36 [.33, .47]**	**.39 [.26, .50]**
PE	36	.34 [.01, .59]	.03 [-.29, .34]^{a}	.65 [.39, .80]
FNE	55	-.05 [-.30, .21]^{a}	.19 [-.07, .42]^{a}	.02 [-.24, .27]^{a}
FOMO	36	.52 [.22, .71]	.33 [-.001, .58]^{a}	.42 [.10, .64]
MW: D&S	28	-.01 [-.37, .35]^{a}	.10 [-.27, .44]^{a}	.26 [-.12, .57]^{a}
NEO-PI-C	45	.36 [.07, .58]	.05 [-.24, .32]^{a}	.49 [.21, .67]

Note. '*n*' refers to the number of inter-item relationships in a scale. For example, the PE scale has 9 items, meaning there are 36 relationships in total. Posterior means [and 95% CIs] shown. '^{a}' marks 95% CIs that include 0.

Regressions (H2 & H3)

The entire pattern of results holds with near identical CIs around the posterior regression coefficient estimates, and similar difference distribution CIs (with identical conclusions).

Table B2

Regressions of the Three Scale Versions' IICs on their Respective SSs.

Model	Effect	Estimate	Est. Error	95% CI		$\hat{R}$
				LB	*UB*	
OG$_{IICM}$ ~ OG$_{SSM}$	Intercept	.21	.04	.14	.29	1.00
	Semantic similarities	**.45**	.10	**.26**	**.63**	1.00
	Sigma	.22	.01	.20	.24	1.00
SP$_{IICM}$ ~ SP$_{SSM}$	Intercept	.29	.03	.23	.34	1.00
	Semantic similarities	**.41**	.08	**.26**	**.56**	1.00
	Sigma	.17	.01	.16	.19	1.00
LP$_{IICM}$ ~ LP$_{SSM}$	Intercept	.22	.03	.17	.28	1.00
	Semantic similarities	**.56**	.10	**.38**	**.75**	1.00
	Sigma	.10	.01	.18	.22	1.00

Figure B1

Scale Versions' Regression Plots

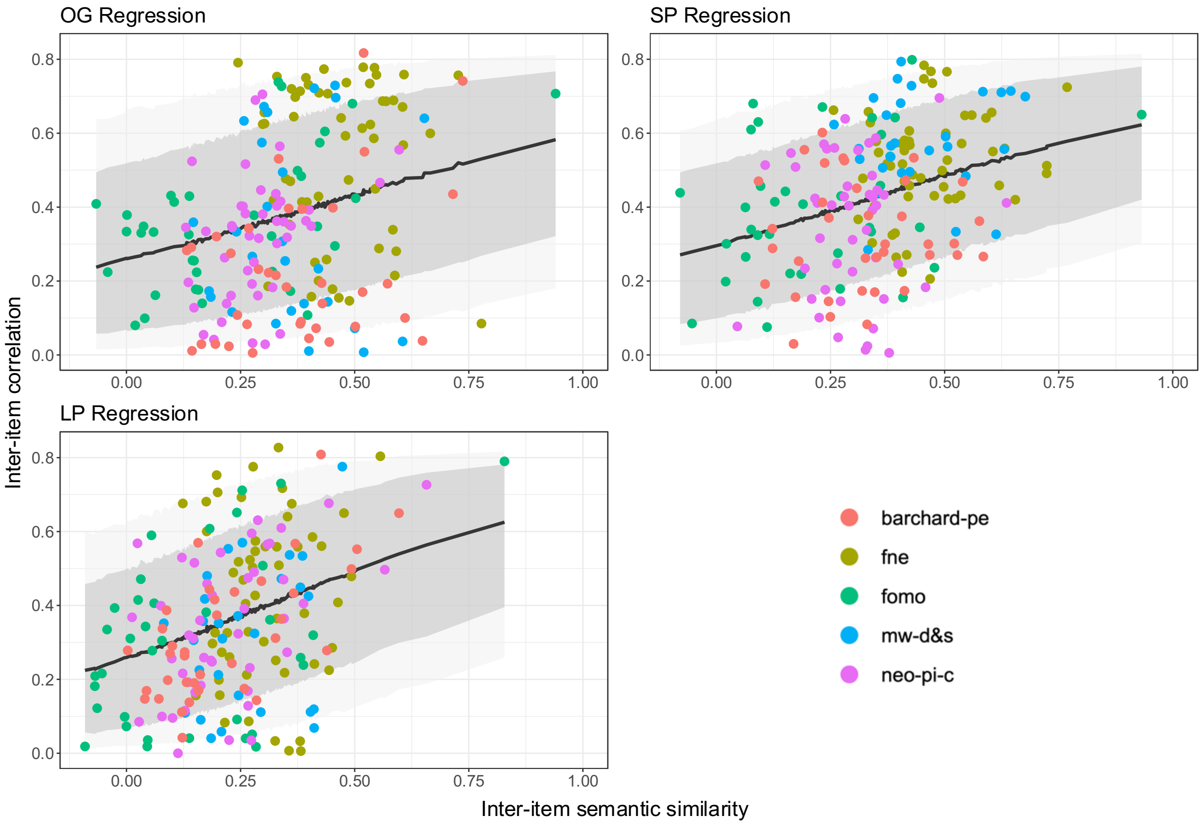


Table B5

Posterior Difference Distributions' 95% CIs

	β_{OG}	β_{SP}	β_{LP}
β_{OG}	–		
β_{SP}	[-.18, .29]	–	
β_{LP}	[-.38, .15]	[-.34, .10]	–

Note. Read column then row – the posterior β_{OG} and β_{SP} difference distribution's 95% CI lies at [-.18, .29].