Received: 31 August 2021 | Revised: 26 July 2022 | Accepted: 17 August 2022

DOI: 10.1111/peps.12543

PERSONNEL PSYCHOLOGY WILEY

ORIGINAL ARTICLE

The AI-IP: Minimizing the guesswork of personality scale item development through artificial intelligence

Ivan Hernandez[1] | Weiwen Nie[2]

[1]Department of Psychology, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, USA

[2]Hogan Assessment Systems Inc., Tulsa, Oklahoma, USA

Correspondence
Ivan Hernandez, Department of Psychology, Virginia Polytechnic Institute and State University, Blacksburg, VA, 24060, USA.
Email: ivanhernandez@vt.edu

Abstract

We propose a framework for integrating various modern natural language processing (NLP) models to assist researchers with developing valid psychological scales. Transformer-based deep neural networks offer state-of-the-art performance on various natural language tasks. This project adapts the transformer model GPT-2 to learn the structure of personality items, and generate the largest openly available pool of personality items, consisting of one million new items. We then use that artificial intelligence-based item pool (AI-IP) to provide a subset of potential scale items for measuring a desired construct. To better recommend construct-related items, we train a paired neural network-based classification BERT model to predict the observed correlation between personality items using only their text. We also demonstrate how zero-shot models can help balance desired content domains within the scale. In combination with the AI-IP, these models narrow the large item pool to items most correlated with a set of initial items. We demonstrate the ability of this multimodel framework to develop longer cohesive scales from a small set of construct-relevant items. We found reliability, validity, and fit equivalent for AI-assisted scales compared to

scales developed and optimized by traditional methods. By leveraging neural networks' ability to generate text relevant to a given topic and infer semantic similarity, this project demonstrates how to support creative and open-ended elements of the scale development process to increase the likelihood of one's initial scale being valid, and minimize the need to modify and revalidate the scale.

KEYWORDS
artificial intelligence, big data, machine learning, personality, personality assessment, technology

1 | INTRODUCTION

Scale development involves many creative, open-ended tasks, including creating concise item text, considering item intercorrelations, and assessing content coverage (Johnson, 2001). Natural language processing (NLP) deep neural networks models offer state-of-the-art performance on a variety of creative and inferential tasks including text generation (Brown et al., 2020), outcome prediction (Raffel et al., 2020), and topic inference (Larochelle et al., 2008). We argue that these abilities, which can synthesize data from millions of text examples into new and/or accurate insights, address the limitations of traditional scale development. Specifically, our approach suggests leveraging the ability of artificial intelligence[1] to (a) learn the common semantic structure of personality items, (b) predict observed associations between psychological constructs, and (c) understand the semantic similarity between sentences. Combining these abilities allows researchers to create comprehensive scales using only a small set of initial items. Specifically, researchers can receive suggestions for additional items that are likely to correlate (both positively and negatively) with all the provided examples in a more automated and reproducible way.

1.1 | Traditional scale development and limitations

Researchers developing personality scales follow a standard series of steps (Hinkin, 1995). The first step requires (1) defining the construct of interest to specify the necessary aspects that the construct should represent, (2) consulting subject matter experts (SMEs) to create a pool of initial items based on that definition. From that initial pool, (3) researchers winnow them to a selection of items that serve as the first draft of the scale. This narrowing requires selecting a sufficient number of items (to obtain acceptable reliability) that cover all content domains of the construct (to promote content validity), and that all inter-relate (to promote unidimensionality). Researchers (4) administer the scale to a large population, and then (5) examine the psychometric properties of the scale. Although there is no substitute for having strong proper construct definitions of constructs, this paper addresses the limitations of the second and third item generation steps. These steps require subjective assessments from SMEs to use their creativity, intuition, and tacit knowledge to generate items matching those definitions. These requirements are problematic because, ideally, the scientific process should be consistent, data-driven, and equally accessible.

SMEs can greatly expand on an item pool from its initial set, but because they rely on their personal experience and perspectives to self-generate items, there is subjectivity, inconsistency, and bias and error in the item creation process (Rush et al., 2016). These items must intercorrelate enough to obtain an acceptable level of unidimensionality in the confirmatory factor analysis and obtain an acceptable level of internal consistency. Therefore, SMEs have to either

mentally approximate these correlations or leave the likelihood of validating unknown until the analysis. In addition to suggesting items with high intercorrelations, researchers who desire sufficient content validity must make sure that the items cover all theoretically meaningful facets (Nunnally & Bernstein, 1994). Depending on the nomological network of the construct, researchers may need to include more items to obtain acceptable levels of reliability. For example, imagine a three-item scale, with each item measuring one of the three theoretically described content areas. If the facets were all correlated at $r = .4$, then, even if the items were parallel, the internal consistency would be .66. With twice as many items, the internal consistency reaches .80. Although shorter scales are more desirable, all else equal, the constraints of examining certain constructs may require a researcher to create additional scales, which will be limited by the creativity of the SME. Therefore, because of limitations in creativity, imprecise item estimations, and the need to establish content coverage, humans have limitations that decrease the effectiveness of scale generation. Because of their expertise, SMEs can be costly to acquire, if the researcher is not an SME or requires multiple. Some researchers may not have the resources or network to obtain these SMEs, lowering the equity of the scale construction process. We suggest addressing this by replicating the insights of an SME with the ability to generate possible items and accurately discern which items are relevant to a given construct.

1.2 | Automated item generation

Automated item generation (AIG) addresses the limitations of human item generation by relying on computational innovations to produce high-quality and large-quantity test items (Gierl & Lai, 2013). Historically, AIG relied either on generating clones of narrowly defined item types such as those found in language-free intelligence tests (e.g., Raven's progressive matrices) or used an analysis of cognitive task components of certain items and derived schemata to produce new item instances (e.g., Bejar et al., 2003; Embretson, 1999; Embretson & Yang, 2006; Gierl & Lai, 2013).

Unfortunately, most AIG approaches only apply to items that are procedural and lack complex semantic information. For example, generating geometry items for a standardized exam only requires changing the values of the figures. For AIG personality assessments, the variation must occur on the item text, whose terms are not as easily substituted, because the new term may lack relevance in the original context. Although some approaches use word replacement by finding approximate synonyms, this approach cannot generate novel grammatical structures, and the new statements would highly overlap with the original construct, limiting the diversity of constructs examinable.

Another approach to AIG is developing a semantic schema for an item, similar to how a "Mad Libs" template works. This approach, however, would have difficulty applying to personality items because of the variety of sentence structures found in personality items. To address these limitations, we propose improving AIG for personality item development using a pipeline of multiple modern deep neural networks to generate a massive, diverse pool of candidate items and identify the candidates most likely to correspond with a set of example items provided by the researcher.

1.3 | Neural network approaches to natural language processing

Neural networks are a type of machine learning method that apply a sequence of mathematical transformations to take input data and convert it into a desired output prediction. Neural networks are useful for solving various NLP problems including classifying text, dimension reduction, and text generations. By increasing the complexity of the neural network's layers and the functions of neurons within those layers (i.e., "Deep Learning"), researchers can solve more complex problems. Neural networks' applicability toward natural language problems and strong prediction ability can potentially address the prior limitations of AIG.

The most related attempt at generating personality items used Recurrent Neural Networks (von Davier, 2018). Recurrent Neural Networks, in contrast to simpler feed-forward networks, contain feedback loops allowing them

to better process sequentially dependent data, such as generating text, where a prediction can be sent back to the network as an input for the next prediction. Unfortunately, the recurrent neural approach is problematic and limited for many reasons. Recurrent neural networks require a large sample of items from the construct of interest. They will produce both sensical and nonsensical statements, requiring a large degree of curation. Further, the prior approach used a character level RNN, meaning it only predicts the next character in sequence compared to the next word, lacking a high-level grasp of language, beyond orthographic patterns. In addition, there is no indication of what construct the generated item represents. RNNs are also less effective as the length of text increases due to the "vanishing/exploding gradient problem," where each successive input's gradient, which is used to update the network's weight, becomes exponentially small/big as it travels to the prior timesteps. This limitation prevents RNNs from learning from long dependencies, such as long sentences. While long short-term memory networks, an advanced type of RNN, can solve the vanishing gradient problem, they contain very long gradient paths, making them extremely computationally intensive. Because of these limitations, RNNs have largely fallen out of favor in NLP. While an RNN model to generate personality items could not distinguish which generated items belonged to a given construct or their overall validity, nor could it consistently provide grammatically correct items, it illustrates that neural networks have capabilities that are relevant to scale development.

1.4 | Transformer models as a solution to assistant scale development

In contrast to the prior used recurrent neural networks, transformer models are the current state of the art for learning from natural language text sequences. Transformers offer several advantages over the prior dominant architecture: recurrent neural networks. Specifically, transformers include a self-attention mechanism that allows the network to remember previous word tokens in the sequence, which can be computed in parallel on the entire sequence. This simultaneous processing of the words in a statement speeds up training and inference.

1.4.1 | Proposed neural network approach to automated item generation

We propose a framework that leverages the improved abilities of transformer networks to solve the prior mentioned issues. Specifically, the proposed framework accomplishes the following aims that prior approaches could not:

1. **Aim 1**: Generate text that is largely equivalent to the structure of human-generated items, with high levels of linguistic acceptability
2. **Aim 2**: Recommend items to a user that are highly related to all other items, especially those the researcher believes best reflect the domain of interest
3. **Aim 3**: Determine whether items cover the broad content areas
4. **Aim 4**: Obtain similar psychometric properties as scales that have already passed the traditional scale development process

Accomplishing these aims can minimize the amount of specialized human labor required. The generation and item assessment aspects play similar roles as SMEs. This method ultimately provides data-driven, educated hints to supplement the expertise of the primary researcher. As long as the researcher has a few focused examples of the construct of interest, the system can quickly expand on those and provide additional items.

Currently, SMEs and researchers must guess whether their self-generated items will validate and form a cohesive scale. Because the method provides items that are all inter-related with the initial items, the scale should be more likely to validate upon the first administration. Alternatively, to avoid having to re-validate, researchers using traditional scale development may have included a surplus of items. These additional items would be more taxing to the respondents, providing potentially shorter administration times.

The recommendation process provides specific quantitative metrics of quality (e.g., estimated inter-item correlation, content domain probability) alongside each suggested item. Therefore, researchers can offer greater transparency for how they selected their items and use a consistent process across various studies.

1.4.2 | Aim 1: Generating realistic items through generative transformer models

The most performant generative transformer models are from the OpenAI group, called "Generative Pre-trained Transformer" or simply "GPT." GPT learns to predict the next word in a sentence/paragraph based on the previous words. It can generate long strings of text by constantly predicting the subsequent word. The OpenAI group demonstrated how to achieve more accurate text generation for a given domain by pretraining a language model on a diverse corpus of unlabeled text, and then fine-tuning the model on a specific task. This GPT framework, therefore, allows text to be generated using a common underlying language model that has a general, task-agnostic understanding of grammar and word sequences. This general, task-agnostic model outperforms state-of-the-art models that used architectures specifically crafted for each task in 9 out of the 12 tasks studied.

GPT was trained with a causal language modeling (CLM) objective making it suited for predicting the next token in a sequence. The CLM objective takes a large corpus of text (e.g., Wikipedia) and breaks it down into smaller sentences of size k. Researchers hold out the last word, and provide the model the first $k-1$ words of the sentence. The model then tries to predict what the held-out word was, given those initial starting words. Thus, the model learns how a series of words "cause" another word to occur. This type of training allows GPT to generate syntactically coherent text. Further improving upon the original GPT model, Open AI developed the GPT-2 and GPT-3 models, though the latter version is not publicly available.

We propose fine-tuning a deep neural network originally designed to create English prose (GPT-2-XL; Radford et al., 2019) to learn the semantic patterns in self-report items. We expect:

> **H1**: *A generative neural network can create personality items that have similar semantic properties to human-created items, including (a) orthographic structure, (b) linguistic acceptability, and (c) indistinguishability by humans, with the benefit of (d) greater lexical diversity.*

1.4.3 | Aim 2: Recommending correlated items through classification models

Another popular task in NLP is "text classification," where a model is provided with a sample of text and must determine its association with a ground-truth value of a construct. Sentiment analysis is an example of a classification task. In sentiment analysis, a model must determine how positive or negative the tone is in a sample of text (expressed using a probability score from 0 to 1). The most performant type of text classification models is extensions of the Bidirectional Encoding Representations from Transformers (BERT) introduced by Google (Vaswani et al., 2017). Similar to GPT-2, BERT provided a large general-purpose neural network that modeled natural language by understanding co-occurrences in text. Specifically, BERT used a "Masked Language Modeling" task to understand the general structure of a language. In the Masked Language Modeling task, a large corpus (e.g., Wikipedia) is provided to the model. Some of the words are "masked" and hidden from the model. The model must determine what the masked word is based on the surrounding unmasked words. This task teaches a model to learn what words are substitutable with each other, given the surrounding context. In contrast to CLM, which lacks later context, this training process facilitates performance on tasks that require encoding the entire meaning of a sentence.

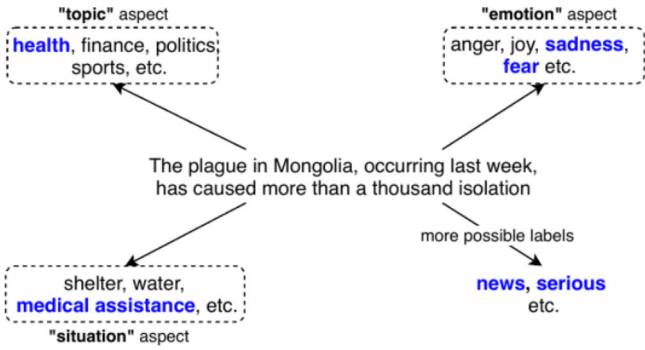

FIGURE 1 Zero-shot classification *models and content analysis. Note*: Illustration of how a zero-shot classification model can be provided with a given text (describing the plague in Mongolia), and be provided list of potential themes (e.g., topics, emotions, situations, tone), and then classify what theme the text best matches.

1.4.4 | Application of BERT models to item generation

We propose using the BERT architecture to infer how correlated two personality items are. Ideally, the BERT model would produce an output embedding for an item, where highly correlated items will receive similar embeddings from BERT. One way to accomplish this goal is using a "paired" framework, where two versions of the same BERT model are each provided one of the statements, and they each produce a fixed length embedding containing the latent representation of sentence's semantic content. The cosine similarity, which ranges from −1 to +1, between those embeddings is computed and compared to the observed correlation between those statements. The mean squared error between the embedding similarity and ground-truth correlation determines how much to adjust the weights in the BERT model. By providing a large number of statement pairs with their observed correlation, the model's weights eventually adjust so that the cosine distance between two statements' embeddings can accurately approximate the correlation between the statements.

> **H2**: *A paired-transformer neural network can estimate, above chance levels, the empirical correlation of personality items of item pairs the model has never encountered. This performance will be reflected in both its (a) mean absolute error, and (b) correlation with the actual ratings, which will (c) supersede the predictions of psychological researchers.*

1.4.5 | Aim 3: Assessing content coverage through zero-shot classifiers

A newer area of research in neural networks focuses on classification models that do not require any prior data to classify text to specific content domains. Historically, if a researcher wanted to create a model to classify a news article as "politics," "economy," or "lifestyle," the researcher would need to amass a large collection of articles from each domain and then use supervised learning to train the classification model to approximate those ground-truth outcomes. Zero-shot classifiers allow a researcher to simply specify the possible content domains. They then score the text's relevancy to each (Figure 1). Zero-shot methods work by associating observed and nonobserved classes through some form of auxiliary information, which encodes observable distinguishing properties of objects (e.g., a semantic embedding of topic labels that can be compared to the semantic embedding of an input text). These models are useful for problems involving predicting topics of text, especially when the content domains (topics) are not the same across user purposes and when there is limited data for the domains. We believe these models can facilitate content validation.

One challenge for applying machine learning to content validation is the lack of sample size coupled with the diversity of possible content labels for any project. Supervised learning would be difficult because a researcher cannot obtain hundreds of items that belong to a single content domain. If they had access to hundreds of items belong-

ing to that domain, they would already have enough items for a scale. In addition, each new project might invoke different content domains, which would require retraining the content classification model. Zero-shot classifiers solve these limitations because they can assign text to a constrained list of topics, without requiring additional training data. Although the zero-shot classification methods can potentially evaluate how well text aligns with potential content domains, a researcher must still define those potential content domains, similar to the traditional process of providing SMEs with content domains to evaluate. Therefore, we believe that zero-shot classifiers can serve a similar role as SMEs in providing insight into how well a given item aligns with specific content domains.

> **H3**: *Zero-shot classifiers can assist with content validation by offering insight into which content domain a personality item belongs to from a list of options. This performance will be reflected in its (a) overall accuracy, (b) precision, and (c) recall of its predictions compared to known content labels.*

1.4.6 | Aim 4: Complete pipeline for AI-assisted item recommendation

We propose combining all methods into a single system that can suggest additional items. This system first requires fine-tuning a generative text model to learn common semantic patterns within personality items and then generating a large (1000,000) item pool of candidate items. It also requires developing a paired transformer model to accurately predict the correlation between item pairs. After developing the item pool and training a correlation prediction model, the system can receive a small sample (e.g., 3) of potential scale items that reflect a construct of interest from a researcher. The system first compares the semantic similarity between those items and the AI-generated item pool based on their Universal Sentence Encoder embeddings. The top candidates then have their correlation with each of the seed items predicted. The candidates with the highest and lowest correlations are then provided to the researcher as a collection of suggested positively scored and negative-scores items. To assist with selecting the final items, the researcher could consult a zero-shot model by specifying the content domains the scale needs to cover and infer the proportion of content covered by the items. We illustrate the framework below (Figure 2) and elaborate on the specific steps in study 4.

The current research project seeks to evaluate the effectiveness of the subcomponents of the framework and demonstrate the effectiveness of the complete proposed framework. We define effectiveness as creating scales for a particular construct of comparable or superior quality to those created by traditional methods. We hypothesize,

> **H4**: *Personality scales developed by the proposed item generation system will demonstrate similar psychometric properties as a human-generated scale that measures the same construct. This similarity will be reflected in the scales' (a) internal consistency, (b) composite scores, (c) and factor structure.*

2 | STUDY 1: ARTIFICIAL INTELLIGENCE ITEM POOL GENERATION

2.1 | Method

The data that support the findings of this study are openly available in the paper's Open Science Framework Repository at https://osf.io/6wznm. All analyses, trained models, and output files are also hosted at the repository.

We fine-tuned the largest freely available version of GPT-2(GPT-2-XL)[2], to learn the semantic patterns in self-report items. To adapt GPT-2 to understand personality scale items, we provided the model with all 3320 items from the largest item repository, the international personality item pool (IPIP) (Goldberg et al., 2006). After fine-tuning GPT-2 for five iterations across the entire dataset, we had the model generate 1,000,000 new items. Because of the large size of GPT-2-XL, we did not explore finding the optimal set of model settings (hyperparameter tuning), as a single

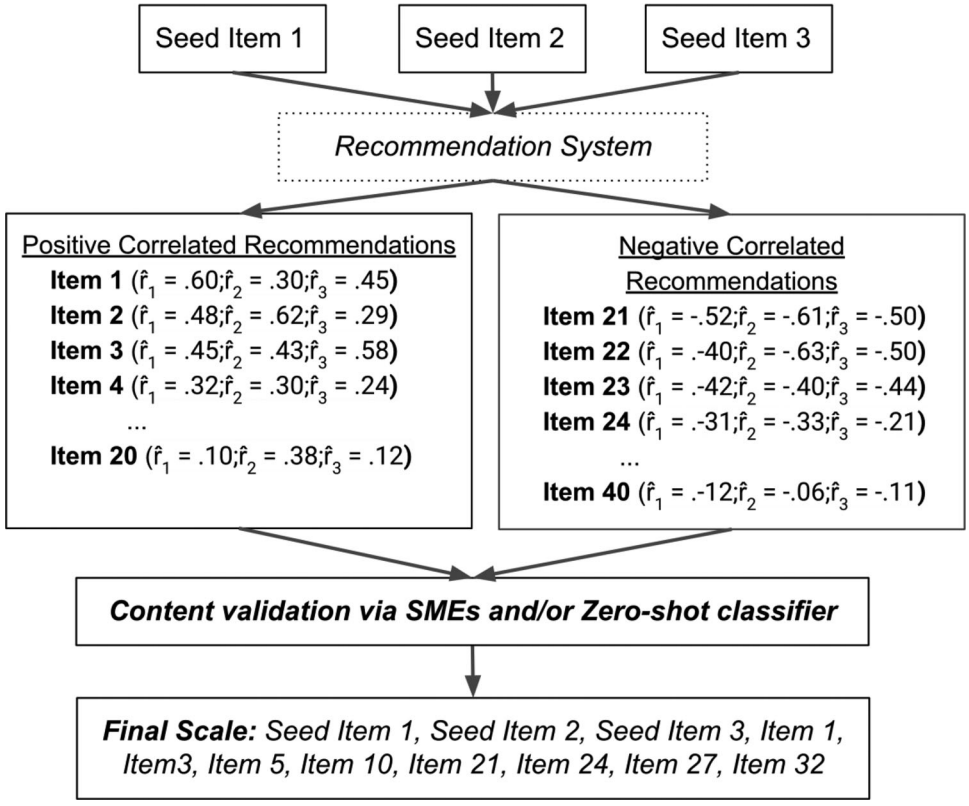

FIGURE 2 Overview of proposed artificial intelligence-assisted scale construction system. *Note*: The figure illustrates how to facilitate the scale construction process if a researcher has an initial set of items related to the construct of interest. These positively correlated seed items are provided to the AI system, which provides additional positive and negative coded construct-related items and their estimated correlation with the items. The researcher selects from those recommendations and consults SMEs or a zero-shot classification system to balance content domains. The selected items are combined with the seed items to form the final scale.

model can take days to train. One hyperparameter commonly varied in neural networks is the number of iterations to pass the training data through the model (i.e., an epoch). Too few, and the model does not have a chance to get accustomed to the target data, reverting to nondomain specific prose. Too many, and the model could overlearn, reproducing only items it has previously seen. We chose to use five epochs as it is aligned with prior GPT-finetuning research and other recommendations, which balances over- and under-learning (Zoph et al., 2016). To counteract the possibility of under-learning, had we trained the model and found that the generated items did not resemble the personality item structure, we would have resumed the model from its last trained state and increased the epochs incrementally until the generated items were deemed acceptable. For our project, five epochs generated items that appeared to capture the desired style.

The GPT-2 model does not inherently create new sentences. Rather, the model is trained with a classification terminal node that predicts how likely each English word would follow an initial sequence. To generate long sentences, generative models are typically combined with a word sampling method. These sampling methods have GPT-2 to make predictions on an input, append the next predicted word to the end of the prior input, and repeat that process until a subsequent "end-of-sequence" token is reached. There are many approaches to creating sentences from generative models. We used "sampling with top-k" approach. This method is useful for explicitly controlling the diversity in the generated items. It works by having the model receive a start input ("beginning of sentence" token), and it then predicts

TABLE 1 Comparison of text between IPIP and AI-generated item pools

Structural metric	IPIP	AI-IP
Average number of words in an item	6.74 (3.38)	5.19 (1.76)
Readability ease	74.88 (30.24)	68.72 (36.60)
Total percentage verbs	22.25%	24.63%
Total percentage nouns	19.47%	22.16%
Total percentage prepositions	10.87%	10.44%
Total percentage adjective	9.00%	10.85%
Total percentage adverbs	6.59%	6.05%
Linguistically acceptable items	99%	93%

Note: Standard deviations are in parentheses. Parts of speech do not have standard deviations because they are the total frequency across all items in the item pool.

the probability of the subsequent word out of all possible words. It then sorts potential next words by probabilities and chooses a word within the top-*J* of those probabilities for the first *I* words, and then within the top *K* for words, after the initial *I* words. Because generation is not as computationally intensive as training a model is, we were able to explore various values for the generation sampling. The values *J* and *K* control item diversity. When they are set to 1 (the lowest value), the model simply predicts the most likely word to follow. When they are set to higher values, words that are less likely are considered as well. Too low, and the items all look the same. Too high and the items could be nonsensical because rare combinations of words are generated. We had different values of *J* and *K* because we wanted the early part of an item to vary more, as this usually contains the central theme of the item (i.e., the construct), and we wanted the later part of an item to be more consistent as generative models often tend to ramble or lose coherency. The value, *I*, serves as a threshold for when *J* and *K* will have their effect. It defines what is considered the early part of the item. We varied the values of *J*, *I*, and *K* to try to create items that seemed sensible, but also diverse. Using simple visual inspection, we eventually settled on having the model generate an item by predicting the next word using one selected from the top-200 (*J*) most likely probabilities for the first 5 (*I*) words, and then within the top-20 (*K*) for words after the first five.

2.2 | Results

We examined the generated item quality by comparing their nonredundancy, diversity of words, and similarity in structure with a human-generated item pool. To examine nonredundancy, we compared the number of identical items between pools. The artificial intelligence-based item pool (AI-IP) contains one million unique items. We first lowercased and removed all punctuation from both item pools. We then compared the exact match between pools and found that 993,714 AI-IP items were unique statements not found in the IPIP. The AI-IP contains 35,426 unique words relative to the IPIP's 3010. We lemmatized each statement (e.g., friends → friend, enjoyable → enjoy) to examine the variety of core word stems in each item pool. The AI-IP contains 21,317 unique lemmas, whereas the IPIP contains 2335 unique lemmas. Therefore, the AI-IP provides novel statements and terms beyond those offered by the largest currently available item pool.

We also compared the sentence structure of the AI-IP relative to the original IPIP. We quantified sentence structure in terms of average item length, readability score, percentage of various parts of speech (i.e., nouns, adjectives, prepositions, verbs, and adverbs). For all metrics, we found highly similar structures (Table 1). Both the AI-IP and IPIP had similar item lengths. The AI-IP on average had items that were approximately 1 word shorter (M_{words} = 5.19; *SD*

= 1.76) than the IPIP (M_{words} = 6.74; SD = 3.38), d = 0.58. In common language effect size terms, there is a 65.79% chance that a randomly selected AI-IP item will be shorter than a randomly selected item from the IPIP (Mastrich & Hernandez, 2021). Therefore, although there is a practical difference in item length, it may be desirable to have more concise items, if they demonstrate equivalent validity. The average word length in the AI-IP ($M_{WordLength}$ = 4.88, SD = 2.53) is nearly identical (longer by less than a character on average) than the IPIP ($M_{WordLength}$ = 4.57, SD = 2.41), d = .13. To examine the complexity of the statements themselves, we used the Flesch Reading Ease metric (Flesch, 1948). It ranges from negative infinity to 121.22. Higher scores indicate easier-to-read statements. It calculates the reading ease of a statement using the formula: 206.835 − 1.015 *(average words per sentence) − 84.6 * (average syllables per word). Therefore, using fewer words per sentence and using fewer syllables per word both increase the readability score.

While, on average, the items in the original IPIP are more readable ($M_{readability}$ = 74.88) than the AI-IP ($M_{readability}$ = 68.72), this difference is also minimal compared to the vast variation in readability within each scale (SD_{IPIP} = 30.24; SD_{AIIP} = 36.60), d = 0.18. The effect size suggests a 5% difference in the probability that a randomly selected AI-IP item will be less readable than a randomly selected IPIP item, compared to another randomly selected IPIP item (Mastrich & Hernandez, 2021). From an absolute perspective, the Flesch Reading Ease scoring guides indicate that a readability score of 68.72 is on the margin between "standard" (60–69) and "fairly easy" (70–79), implying that the AI-IP items are, on average, not difficult to read.

The parts of speech were also highly similar. The percentages never varied more than 3%. The rank order of the parts of speech was highly similar. Verbs and nouns were the most common parts of speech in both, occurring at twice the rate of the other parts of speech. The only difference is that adjectives were slightly less common relative to prepositions in the AI-IP, though the difference between the two is less than .5%. Therefore, the grammatical components of the item pools largely converged.

We evaluated how grammatically sensible the items were by applying a neural network classification model developed to assess linguistic acceptability. Linguistic acceptability is a determination of whether a statement follows standard grammatical conventions in English. Features of text required to be acceptable include: correct spelling, subject-verb agreement, complete punctuation, proper capitalization, among others. For evaluating linguistic acceptability, we used the RoBERTa-base-CoLA model (Morris et al., 2020), which obtains 85.04% accuracy on the COLA benchmark, and therefore serves as a reasonable approximation to a human evaluation of acceptability. Approximately 93% of the AI-IP items are linguistically acceptable, compared to 99% of the IPIP items. Therefore, the computed generated items are more likely to suffer from grammatical issues, but the vast majority of items are grammatically correct.

Because humans may be able to capture item peculiarities not reflected in just grammar, we followed-up the computational analysis of linguistic acceptability with an empirical examination of how trained researchers could distinguish the origin of human and computer-derived items, on a randomly selected subset of the item pools. We presented a random selection of 100 items (50 from IPIP and 50 from AI-IP) to eight I–O doctoral students (62.5% holding master's degrees) with an interest in research methodology, and who all had prior professional experience in applied Industrial–Organizational positions. Participants were shown 200 items from the IPIP prior to the task, to illustrate what human-written items look like. Students were told they would be shown 100 items—some that were written by a human and some written by a computer. They were asked to indicate the source of the item (human vs. computer) All students were blind to the hypotheses and were told that the student in the group with the highest accuracy would earn a gift certificate to a local business. The average accuracy of the entire group was 57.07%, which is not statistically distinguishable from the guessing base rate of 50%, $Z(100)$ = 1.4, p = .16). This value is highly similar to the human-detection rates of GPT (52%; Brown et al., 2020). The highest accuracy any rater obtained was 65%, suggesting that there was no overwhelming ability to detect the origin of a statement. These results highlight that the majority of unedited statements are not easily identifiable as originating from a computer, even when respondents are told that some of the statements are written by a computer. The subtlety could also be improved by first applying a linguistic acceptability model when compiling the one million item pool.

2.3 | Study 1 discussion

This first study demonstrated the ability of generative neural networks to write semantically novel personality items. The generated items are slightly shorter, with a greater diversity in length, and comparable in terms of general reading ease and phrasing. In addition, 93% of the items are considered linguistically acceptable. Researchers show no strong ability in detecting the difference between human-generated and computer-generated items. Our next study evaluates the ability of the framework to predict the correlation between item pairs. This subsequent aspect of the framework is necessary to leverage the artificial intelligence-generated item pool by predicting the correlation between a researcher's initial seed items and the massive pool of candidate items.

3 | STUDY 2: ITEM PAIR CORRELATION PREDICTION MODEL

The second unique contribution of the framework is an NLP model that receives a pair of personality items and predicts the observed correlation between those items. This model uses a paired architecture, which optimizes the weights of a neural network, so they produce outputs to statement pairs that approximate the correlation between those statements. In this section, we describe the model training process and report on the model's performance at predicting item–pair correlations extracted from the Open Psychometrics data repository[3].

3.1 | Method

Training a functional NLP model that is based on a deep neural network is an arduous task, which often requires significant strong computational power and/or long waiting time. To overcome these barriers, we employ a transfer learning strategy (Raffel et al., 2020), which uses a model pretrained on a similar problem as the initial weights, serving as a headstart for finding optimal weights for the current problem. For this study, we will implement transfer learning based on the Sentence-BERT model (Reimers & Gurevych, 2019).

We chose the DistilBERT architecture, which is a smaller, approximate version of the full BERT model. The "distillation" process in neural networks trains a smaller version of the model to predict the predictions of the larger model. DistilBERT obtains nearly the same accuracy on natural language tasks as its larger predecessor. Using a smaller model takes less time to train and also offers more compatibility across different user systems, which may have limited resources. Our starting set of network weights came from the DistilBERT model, "distilbert-base-nli-stsb-mean-tokens," which was trained to detect similar sentences on the Semantic Textual Similarity benchmark task, which we chose because predicting correlations between items involves inferring semantic similarity (Sanh et al., 2019).

The paired DistilBERT model trains a model that accepts two text inputs and a ground truth similarity (Figure 3). By providing a series of these examples to the model, with known similarities (i.e., correlations between statements), it optimizes the model's weights so that the model is able to predict the similarity of two texts provided to it later. The specific steps of this paired text training first involve including two identical DistilBERT networks parallel to the neural network, each with the same exact transferred weights. The DistilBERT network used creates an output of 768 dimensions for each 128 possible tokens (reduced from the original 512 for computational simplicity). To obtain a single-dimensional embedding vector of 768 values, the network includes a pooling layer after the final DistilBERT layer, which averages across the 128 input tokens to create a single 768-dimension embedding. A text is passed through one network, and another is passed through that network's "twin." Third, the output of the two pooling layers is given to a new layer, which calculates the cosine similarity. The range of the cosine similarity is naturally scaled from −1 to 1, which matches the natural scaling of correlation coefficients. Because the cosine similarity (equivalent to the normalized Euclidean distance between two vectors) of the two texts is used to calculate the prediction, the order of sentences entered into the model does not change the prediction outcome. For this study, we define the loss function

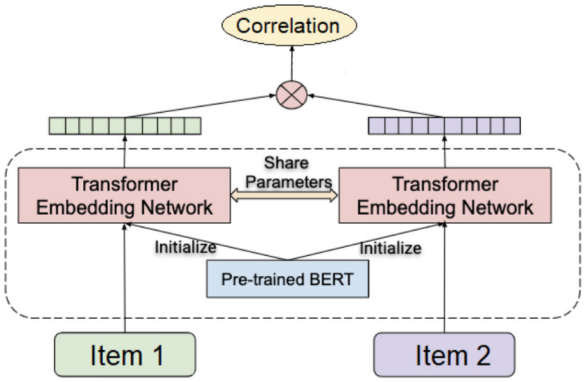

FIGURE 3 Visualization of the paired BERT model to predict the correlation between two items. *Note*: The model above shows the neural network architecture of a paired BERT model that accepts two input texts and adjusts the weights within the Transformer network so that their embeddings from that network have a cosine similarity equal to their correlation.

of this neural network by taking the mean squared difference between the interitem correlation and the cosine similarity of the two items. The error is used to update the BERT model's weights in the same exact way, so both networks always have the same exact weight values.

To train the NLP model for this study, we used all the publicly available personality datasets from the Open-Source Psychometrics Project[3] and the IPIP dataset from the Eugene-Springfield Community Sample (Goldberg, 2018). We chose these datasets because they are part of open-source projects that explicitly endorse re-use, offer large sample sizes (all $N > 1000$), and come from international populations. All the inter-item correlations were identified with the correlation matrices in these datasets. We identified a total of 4067 unique item statements with 3,452,812 inter-item correlations across all datasets. The total number of inter-item correlations is smaller than the maximum possible (4067 choose 2 = 8,268,211) because not every pair of statements was part of the same study. We then fed both the item text pairs and the inter-item correlations that related to these items into the neural network to fine-tune the neural network.

The fine-tuning process consists of using a pretrained model in a related domain as the starting point for training the model to new data in the current domain. By fine-tuning, we used a sentence similarity model that had been pretrained to identify how closely related two sentences are. The outputs of the model were already scaled from −1 to +1, which conveniently is the same range as a Pearson's correlation coefficient. Therefore, we made no changes to the architecture underlying the model. The fine-tuning process merely re-estimates the weights within the network to optimize predictions in the current domain. We trained the model for 10 epochs, which was an arbitrary choice. Typically, models are trained for fewer epochs, but we were concerned about potentially undertraining the model, so we opted to have a higher number of epochs, and monitored the cross-validated loss to verify that we were ending too soon. The cross-validated correlations showed stability (neither increased or decreased more than .01) at around the 5th epoch, and therefore it is likely that our results would not change if we further increased the number of epochs, suggesting that five epochs would have been sufficient. To evaluate the quality of the model at predicting the correlation between items, we held out a percentage (10%) of inter-item correlations from the model.

3.2 | Results

3.2.1 | Prediction performance

We compared the predicted correlation between 345,282 item pairs (10% data) that were held out during the training phase. The cross-validated correlation was $r = .96$, $p < .0001$). The cross-validated mean absolute error of predictions was equal to .02 ($SD = 0.018$, 95% CI = [−.015, .055]). This error was significantly lower than the baseline error

(0.081) expected by guessing the mean correlation ($r = .018$). To evaluate the amount of error that could be expected in practice, we examined the list of all absolute errors and examined the percentiles at various cutoffs. At the 95th percentile, the absolute error was 0.055. Thus 95% of the time, the predicted correlation was not off by more than .055 from its true value. At the 99th percentile, the error was 0.08 and at the 99.9th percentile, the error was 0.14. Therefore, it is extremely unlikely that the model will provide a predicted correlation more than .15 off its true value. We also compared the predictions to estimates made by graduate researchers (see study 1 for sample description). The mean absolute error of humans was .10 ($SD = 0.02$, 95% CI = [.09, .12]), and the average correlation between their predicted and actual values was .38 ($SD = 0.09$; 95 CI = [.31, .46]). Therefore, the model at its 99th percentile or error, substantially outperforms how well even researchers can generally estimate the item correlations.

3.3 | Study 2 discussion

Our second study demonstrated the ability of transformer neural networks to predict the correlation between pairs of text. Because we calculated the model's error on a hold-out sample, we express confidence in the model's ability to predict the correlation of new pairs of text. Validating this component justifies its inclusion in the framework, as it winnows the list of candidate items down to the ones expected to have the highest correlation with the user-submitted items. Further, it offers researchers a predicted correlation between each recommendation and each submitted item. These correlations can assist the choice of items to maximize validity. Our next study evaluates the ability of the framework to generate scales that have comparable psychometric properties to existing validated scales that have successfully completed the iterative development process.

4 | STUDY 3: CONTENT VALIDATION MODEL

The third study examined the effectiveness of zero-shot classification models to facilitate the content validation process by accurately inferring the content domain of an item when provided a selection of domains. For this study, we selected a multitude of pretrained zero-shot classification models and applied them to a dataset with personality items labeled to specific constructs. We then evaluated how accurately each model could select the correct construct, as well as the aggregation of the models.

4.1 | Zero-shot classification models

We selected a variety of zero-shot classification models available from the HuggingFace transformers repository. We tried to obtain a variety of models, pretrained on different datasets, to have a diversity in the perspectives offered. Because we had no a priori predictions of which models would be sufficient, we chose to obtain a larger group of models for exploring what are the most optimal for inferring a construct from an item. We chose the following eight models, which provide insight into how various architectures and pretraining tasks, perform at inferring the correct content domain: (a) Bart-Large-MNLI, (b) Deberta-Large-MNLI-Zero-CLS, (c) GTP-T5-Large, (d) STSB-RoBERTa-Large, (e) All-MiniLM-L12-v2, (f) Sentence-T5-Base, (g) All-MPNET-Base-v2(h) NLI-DeBERTA-v3. There are many models to choose from, so the subsequent conclusions should not be interpreted as how all zero-shot classifiers will perform, but rather how well existing models *can* perform, and their potential utility as even more accurate models are developed.

4.2 | Item–construct assignment data

To evaluate the zero-shot models' effectiveness at correctly classifying an item's content domain, we used the item–construct map provided by IPIP.org. This data contains all of the 3320 items in the IPIP and the specific domain(s) associated with each item. In total, there are 3805 items–construct maps, as some items are associated with more

than one domain (e.g., risk-taking, thrill-seeking), which is beneficial because we would not necessarily want to assume that an item can only correspond to a single facet out of all possible facets. Similar to an SME who has to evaluate which of a limited number of domains an item belongs to, we presented the model with other constructs sampled from the list of all constructs. Specifically, for every row in the dataset, we retained the true construct, and randomly sampled four other constructs from the list that were not also listed with the item. We chose a total of five content domains because many popular scales have similar numbers of facets (IPIP has 6 facets per factor, HEXACO has 4 facets per factor; Emotional Intelligence contains four facets within the overarching factor; Self-Efficacy has four content facets, etc.). When provided the original item to the model and then recorded the probabilities assigned to each of the five potential labels. The label with the highest probability was considered the model's predicted domain.

4.3 | Results

We evaluated each zero-shot classifier's predictions against the ground-truth construct labels for each item. The accuracies ranged from .57 to .70, and were 64.31% accurate ($SD = 0.05$), on average. The best performing model was the Sentence-t5-base, which is especially suited for sentence similarity tasks (Ni et al., 2021), whereas the worst performing model was the all-MiniLM-L12-v2, which was trained on a variety of data inference tasks (Reimers & Gurevych, 2019). Content validation often aggregates the assessments of different SMEs to make a final determination of content relevance. We combined the model's predictions and evaluated if the inferences improved using the label inferred by the plurality of the models. In this aggregation, each model's most likely content label received one vote, and we selected the candidate with the highest number of votes. Selecting the content domain that received the most predictions led to the highest accuracy (73.96%), which is more than three times the 20% base rate expectation. If we use a higher decision threshold, then we can obtain even higher accuracy, at the expense of missing some construct-relevant items. We found that by setting the decision threshold at 60% of models must agree, the accuracy increases to 84.29%, but will not make a definitive prediction for 25% of items. Requiring 80% of the models to agree, increases the accuracy to 92%, but the model will not reach consensus for 58% of items it evaluates. Therefore, researchers with many items to choose from may desire to set the consensus threshold to 80% to increase the probability that the items selected actually match their classified domain.

4.4 | Study 3 discussion

Our third study demonstrated that zero-shot models can perform above guessing levels when inferring the content domain of an item. The models were not perfect, missing the correct construct 30% of the time. However, we believe that these models, if used in tandem with an SME who is similarly accurate, can offer additional insights to minimize the errors made by a human alone. That is, if a second SME cannot be acquired, and if the model's and first SME's errors are independent from each other, then, when both agree, the accuracy should approach 91% ($1 - .30 \times .30$). The error should improve even further if using a higher decision threshold for the zero-shot model, at the expense of falsely rejecting relevant items.

5 | STUDY 4: EMPIRICAL VALIDATION OF SCALES CREATED WITH AI-RECOMMENDATIONS

Based on the results of studies 1–3, we found evidence validating the individual elements of the item generation framework. To validate the effectiveness of the proposed generation framework, we constructed scales using the framework that had validated counterparts whose psychometric properties could be compared to the generated

scales. Specifically, for study 4, we provided three "seed" items for a given facet and generated an equivalent length scale of completely generated items that correlated with those "seed" items. We then surveyed participants who completed both the original and AI-generated scales. By examining the reliability, unidimensionality, and convergence of the scales, we assess how well the proposed framework can generate scales compared to traditional methods.

5.1 | Method

5.1.1 | Artificial intelligence-generated scales

We applied the proposed framework described in the introduction to generate inventories using a set of items from the various validated scales. These scales were the Big Five Inventory 2 (Soto & John, 2017), The Self-Efficacy Scale (Sherer et al., 1982), and the Curiosity Scale (Kashdan et al., 2018). We chose these scales as a representation of a broad range of constructs, with some like the Five Factors, which are in the IPIP, and others like the Curiosity Scale that are not mapped to the IPIP.

Using these scales, we then submitted the three highest loading items ("seeds") for each facet from the scale to a universal sentence encoding model that selects the top .5% most semantically similar AI-IP items from the list of one million items. These 5000 items ("candidates") are then submitted to the correlation prediction model, which predicts the correlation between each seed and each of the 5000 candidates. We then calculate the average correlation for each candidate across all three seed items. The model returns the top 25 most positively correlated items and top 25 most negative correlated items. We then selected as many positively and negatively coded items from the 50 recommended items as the original scale had. Because the highest correlated suggestions may overlap with each, we also prioritized selecting items that were diverse in content, similar to relying on an SME to choose content. Researchers could also use a zero-shot classifier to select items that were classified as belonging to specific content domains, balancing the domains chosen. To counteract the potential inflation of item loadings that would be observed by using a prevalidated item from an existing scale, we never included the initial seed items in the final scale. In practice, researchers would want to include their seed items, which likely greatly maximize obtaining a coherent factor structure. For the initial seed items and a sample of suggested items chosen from those seed items see Appendices A–C. The comprehensive set of guidelines we followed (and encourage others to follow) for obtaining recommendations from the automated system can be found in the supplemental material (S1) and listed on the paper's OSF repository: https://osf.io/6wznm.

5.1.2 | Procedure

Participants ($N = 377$) accessed the survey via an online Qualtrics link obtained after signing up for the study on the Subject pool participant sign-up page. The online Qualtrics survey consists of an online consent form. After reading and agreeing to the consent and meeting the screening requirement, participants will be directed to the next section containing the actual survey questions. Participants saw each scale on its own page. The page never mentioned which scales were being shown and no participant was told that some of the scales were generated by a computer. Each scale consisted of Likert-style questions asking about the cognitive and behavior tendency of the participants. The order of the scales was randomized and all participants eventually saw the BFI, Self-Efficacy, and the Curiosity scale as well as their AI-generated counterparts on six different pages. After completion of the personality scales, participants completed multiple types of attention and effort check questions, including psychological antonyms, asking participants to select a specific answer, asking about highly improbable events, and self-reported inattention (Aust et al., 2013; Johnson, 2005). Last, participants completed basic demographic questions. We removed any participant who failed any of the different attention checks for a final sample size of 277 respondents.

TABLE 2 Descriptive statistics of original and artificial intelligence-generated scales

Scale	Facet	Original			AI-generated		
Big Five		**Mean**	**SD**	**CI**	**Mean**	**SD**	**CI**
	Agreeableness	3.86	0.55	(3.79, 3.92)	4.28	0.59	(4.21, 4.35)
	Conscientiousness	3.72	0.68	(3.64, 3.80)	3.61	0.67	(3.53, 3.69)
	Extraversion	3.39	0.69	(3.30, 4.47)	3.56	0.76	(3.46, 3.64)
	Neuroticism	2.97	0.81	(2.87, 3.06)	3.09	0.79	(2.99, 3.18)
	Openness	3.66	0.66	(3.58, 3.74)	3.45	0.71	(3.37, 3.54)
Self-Efficacy							
	General	3.52	0.50	(3.45, 3.59)	3.54	0.50	(3.49, 3.60)
	Social	3.20	0.66	(3.12, 3.27)	3.75	0.64	(3.68, 3.82)
Curiosity							
	Joyous Exploration	5.13	1.12	(5.00, 5.25)	5.66	1.03	(5.54, 5.78)
	Deprivation Sensitivity	4.54	1.33	(4.38, 4.68)	4.85	1.17	(4.72, 4.99)
	Stress Tolerance	4.22	1.39	(4.06, 4.38)	4.14	1.26	(3.99, 4.29)
	Social Curiosity	5.48	1.02	(4.34, 5.58)	5.90	0.97	(5.79, 6.02)
	Thrill Seeking	3.95	1.36	(3.79, 4.10)	4.02	1.50	(3.84, 4.18)

Note: Values in parentheses are 95% confidence intervals.

5.1.3 | Proposed analysis

We compared the original scales with the artificial intelligence-generated scales using a series of analyses standard in developing equivalent scales. Specifically, we closely followed the procedure of Ehrhart et al. (2009), which provides a comprehensive collection of guidelines and standards for comparing validation metrics between original and alternative scales.

5.2 | Results

5.2.1 | Descriptive statistics comparison

The means and standard deviations for most scales appear to be highly similar (Table 2) with most means within the same scale anchor as the other mean. The average standardized difference between original and AI-generated scales was $d = 0.32$, which is commonly qualitatively described as a "small" difference. Although the practical effect size may be small, these differences do appear to be statistically robust. We found statistically significant differences in means for "Agreeableness," "Social self-efficacy," "Joyous Exploration," "Deprivation Sensitivity," and "Social Curiosity". Interestingly, when the difference is statistically significant, the AI-generated items always had higher means, suggesting that they may be more susceptible to acquiescence biases.

5.2.2 | Reliability comparison

We evaluated the similarity of the reliability coefficients for each subscale. We used Cronbach's alpha to measure internal consistency. When comparing the Cronbach's alpha of the generated scales to the original scales, the generated scales were comparable (Table 3). The majority of the time (91.67%), the generated scales produced Cronbach's

TABLE 3 Comparison of reliability between original and artificial intelligence-generated scales

Scale	Facet	Original α	AI-generated α
Big Five			
	Agreeableness	.77 (.72, .80)	.88 (.85, .90)
	Conscientiousness	.87 (.85, .89)	.85 (.82, .88)
	Extraversion	.85 (.82, .87)	.89 (.87, .91)
	Neuroticism	.90 (.87, .91)	.89 (.86, .90)
	Openness	.84 (.81, .87)	.87 (.85, .89)
Self-Efficacy			
	General	.89 (.87, .91)	.84 (.81, .86)
	Social	.72 (.66, .76)	.74 (.69, .78)
Curiosity			
	Joyous Exploration	.87 (.84, .89)	.91 (.89, .93)
	Deprivation Sensitivity	.84 (.81, .87)	.85 (.81, .87)
	Stress Tolerance	.88 (.86, .90)	.81 (.76, .84)
	Social Curiosity	.81 (.77, .85)	.89 (.86, .91)
	Thrill Seeking	.85 (.82, .88)	.90 (.88, .91)

Note: Values in parentheses are bootstrapped 95% confidence intervals for the alpha coefficient.

alpha coefficients that were statistically indistinguishable from or better than the original scales' alpha. In addition, the generated scales never had alpha coefficients less than .04 of the original scale's alpha coefficients. These comparisons suggest that there is not a strong difference in reliability in artificial intelligence-generated scales and their optimized counterparts.

5.2.3 | Composite score comparison

We compared the composite scores for each of the distinct scales by using the equal-weighted scoring rules from the original scale. We then examined the correlation between participants' composite scores from the original scale and the AI-Generated Scale (Table 4). All correlations were statistically significant at the .05 alpha level. On average, the correlation between the composite scores from the generated scale and the actual scales was .72, which surpasses traditional reliability thresholds. However, these correlations showed a fair amount of variability. For the Big Five, whose underlying constructs were captured in the IPIP that were part of the training data, the correlations are higher. However, for the Curiosity Scale, which did not have any equivalent constructs in the IPIP, the correlations for the subfacets are lower, ranging from .39 ("Social Curiosity") to .72 ("Thrill Seeking"). Therefore, the method did not always produce correlations at reliability thresholds. Specifically, for the Curiosity Scale, the dimension of "Social Curiosity" was especially difficult for the method to generate equivalent items.

5.2.4 | Confirmatory factor analysis

This analysis describes how well the scales conform to the theorized multidimensional factor structure. Because we generated scales at the facet level, we submitted the scales to a confirmatory factor analysis, where the items loaded onto the facet scales and the facet scales loaded onto a single higher-order factor. To evaluate the overall model fit,

TABLE 4 Correlation of composite factor scores between original and artificial intelligence-generated scales

Scale	Facet	Correlation coefficient
Big Five		
	Agreeableness	.76*
	Conscientiousness	.85*
	Extraversion	.91*
	Neuroticism	.84*
	Openness	.80*
Self-Efficacy		
	General	.81*
	Social	.75*
Curiosity		
	Joyous Exploration	.64*
	Deprivation Sensitivity	.51*
	Stress Tolerance	.55*
	Social Curiosity	.39*
	Thrill Seeking	.72*

*Statistically significant at $p < .05$.

TABLE 5 Correlation of composite factor scores between original and artificial intelligence-generated scales

Scale	Original RMSEA	AI RMSEA	Original SRMR	AI SRMR
Big Five	.070 (.067, .073)	.067 (.064, .070)	.095	.088
Self-Efficacy	.084 (.077, .091)	.086 (.079, .093)	.076	.078
Curiosity	.090 (.084, .097)	.092 (.086, .099)	.081	.092

we used the (a) Root Mean Square Error of Approximation (RMSEA; Steiger, 1990), and (c) Standardized Root Mean Square Residual (SRMR). The prior cited research emphasizes RMSEA and SRMR because they are not sensitive to certain types of simple structure distortion (Beauducel & Wittmann, 2005) and are fairly robust to varying reliability values of the indicator variables (Browne et al., 2002). Determining model fit is highly discussed and debated, particularly with respect to threshold values implemented for descriptive fit indices (Lance et al., 2006).

The RMSEA values, which have an associated 90% confidence intervals, are statistically indistinguishable between all three human- and AI-generated scales (Table 5). In addition, the SRMR values, which do not have a confidence interval, are all within .01 of each other. These findings highlight the ability of the generated scale items to have statistically indistinguishable psychometric properties as scales generated by traditional methods.

6 | DISCUSSION

In summary, we present a framework that combines different modern NLP transformer models including, GPT, BERT, zero-shot classifiers, to alleviate some of the challenges of scale development. This framework offers data-derived, informed suggestions to help researchers create cohesive scales. Study 1 demonstrated that generative neural network models can produce a greater variety of personality items than the largest existing item pool, while still having

similar structural properties. Study 2 demonstrated that a paired transformer model can closely estimate the correlation between item pairs using only the items' text. Study 3 provided evidence that pretrained zero-shot classifiers can determine an item's content domain. Study 4 showed that the scales developed by this method perform similarity in terms of reliability, composite score creation, and showing the anticipated factor structure. While SMEs are still necessary at various stages, such as defining content domains, and selecting the final itemset, this approach can hopefully increase the ease that those scales are constructed.

Because the framework is intended to facilitate the research pipeline, we provide access to the framework through an interactive website[4]. The modularity of the proposed framework offers researchers the ability to improve upon the various components and adjust the settings (e.g., # of filtered items, # of returned suggestions). To allow researchers to easily make these adjustments, we also offer the combined suggestion pipeline as a Google Colab notebook.[5]

6.1 | Additional benefits

6.1.1 | Large preconstructed item pool

The AI-IP items that we developed in study 1 represent the largest general personality item pool available with one million unique items, compared to the previous largest pool—IPIP—consisting of 3320 items. Therefore, the project offers more candidate items for researchers by three orders of magnitude (10^3 vs. 10^6) who are solely seeking item banks like the IPIP. Because the AI-IP was arbitrarily limited to 1,000,000 items, researchers who desire more item variety could use the model to create even more items. We have created an online, interactive Google Colaboratory notebook,[6] which offers researchers access to the trained GPT model from study 1 as well as the generation functions, to generate more items on demand. Making this code available should benefit researchers, who are studying more niche content areas that may not be able to be captured by the original million.

6.1.2 | Novel model to predict item correlation

While not the main goal of the project, the item prediction offers a model trained to predict how correlated items are on scales. This model contains a representation of the correlation between semantic concepts, which could potentially be useful for anticipating the results of new studies. That is, if researchers are interested in examining the relationships between concepts, but are unsure about how strongly they would be related to each other, this model could suggest the future outcome of unconducted research projects. It could also help with inferring the effect size to enter in an "a priori" statistical power analysis. Because the model's error was still low on held-out data, this possibility is reasonable, though it likely requires additional investigation to examine the limitations of this type of prospective analysis. We have also created an interactive Google Colaboratory notebook to allow researchers to load the trained item-prediction model and submit their own item pairs to facilitate this area of research.[7]

6.2 | Limitations

6.2.1 | Limited range of constructs explored in validation study

The prior provides initial evidence of the framework's efficacy. However, we do not intend for the evaluation to be exhaustive, but rather limited by practical considerations of how many surveys could be completed by participants before experiencing fatigue effects. We do feel that, like all proposed frameworks, further testing is helpful to explore

its boundaries. The results of the validation suggest that the method works best when construct scales captured within the constructs measured by the IPIP, and that novel construct may require more training data.

6.2.2 | Generalizability to additional populations

The original correlation prediction model was trained using online survey data compiled from various countries. This aggregation can lead to greater country-specific error for items moderated by culture. In addition, the items are all in English, and therefore, the model can only provide suggestions for English language scales. This limitation is currently being addressed in neural network research with the development of multilingual models that provide equivalent predictions for semantically equivalent sentences, across a variety of languages (Reimers & Gurevych, 2019).

6.3 | Future directions

6.3.1 | Larger item pools

One future direction for the proposed method is generating more conceptually diverse item pools. The development of the model used scales published on Open Psychometrics, which have been mapped to approximately 248 constructs. However, many other scales exist within the personality literature and may provide greater construct diversity for the existing items. Training the generative model with these additional scales could facilitate a broader range of themes.

6.3.2 | Grammar correction models

Our research found that approximately 8% of the items in the pool were linguistically unacceptable. Rather than use these grammar-checking models retrospectively, future versions of this system might want to apply them prospectively when developing the initial item pool. As grammar correction models improve in computer science, these models can serve as a more effective filter for whether a GPT-generated item is included in the final item pool.

6.3.3 | Exploring further zero-shot models

Another way to further improve the intelligence of the framework would be to examine additional zero-shot classifiers to determine their optimal combination and more accurately infer what content an item represents. In the current framework, we used pretrained zero-shot classifiers that were not optimized toward psychological language. The current results merely demonstrate that zero-shot models have promise in assisting content validation. As better models are discovered, they become more substitutable for how SMEs would evaluate an item's content.

6.3.4 | Computer adaptive testing

The current framework has the potential to apply to domains that require large item pools, with diverse characteristics. Computer adaptive tests are one potential domain where items become increasingly more difficult or easy depending on the person's prior performance. This format allows researchers to more efficiently estimate a person's underlying ability level on a construct, and also prevent item leakage by having respondents encounter different items

than others. If the current system was combined with methods that could predict an item's item response theory parameters (e.g., difficulty, discrimination, guessability), then these parameters could be stored alongside precreated items. When a respondent requires an item with a certain parameter values, the system would search for the item with the closest matching properties. This approach could allow for the precalculation of millions of items, creating more fine-grained options and also avoid question leakage. This approach could be used in tandem with other classification models designed to filter out grammatically incorrect or culturally biased items.

7 | CONCLUSION

In this paper, we validated the different elements of the proposed item generation method and demonstrated its ability to create scales with similar psychometric properties as scales that had been developed and optimized with traditional scale development pipelines. This novel approach combines the latest advances in machine learning to facilitate a typically creative and subjective endeavor. By leveraging these advanced models, we further the scientific goals of quantifying open-ended processes, minimizing uncertainty, reducing guesswork, and enhancing replicability.

DATA AVAILABILITY STATEMENT

The data, models, analyses code, and output that support the findings of this study are openly available in the manuscript's associated Open Science Framework Repository at: https://osf.io/6wznm.

ORCID


Ivan Hernandez https://orcid.org/0000-0002-3141-7525

NOTES

[1] By artificial intelligence, we mean primarily "narrow AI," which refers to neural network-based and machine learning methods that are optimized to solve specific tasks through iterative examples.

[2] As of time of writing this article, the more advanced GPT-3, has a public API, but is not yet open-source, and we favored using an approach whose underlying model was accessible to all scientists).

[3] https://openpsychometrics.org

[4] Details for accessing and using the recommendation system can be found at: https://osf.io/6wznm

[5] https://colab.research.google.com/drive/1OOGjLbPYvyL1yGHoqJmzIgSCk8Vc5k1f

[6] https://colab.research.google.com/drive/1Fzd1PtEVEVfzOBahUvQmfZGlJ3xUyLZC

[7] https://colab.research.google.com/drive/1Y2FyCcyRUxmHLON8nm5B1QB-Eie0Mz8U

REFERENCES

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. Behavior Research Methods, 45(2), 527–535. https://doi.org/10.3758/s13428-012-0265-2

Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(1), 41–75. https://doi.org/10.1207/s15328007sem1201_3

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *The Journal of Technology, Learning and Assessment*, *2*(3) https://ejournals.bc.edu/index.php/jtla/article/view/1663

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … , & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. http://arxiv.org/abs/2005.14165

Browne, M. W., Maccallum, R. C., Kim, C.-T., Andersen, B. L., & Glaser, R. (2002). When fit indices and residuals are incompatible. *Psychological Methods*, *7*(4), 403–421. https://doi.org/10.1037//1082-989X.7.4.403

Ehrhart, M. G., Ehrhart, K. H., Roesch, S. C., Chung-Herrera, B. G., Nadler, K., & Bradshaw, K. (2009). Testing the latent factor structure and construct validity of the Ten-Item Personality Inventory. *Personality and Individual Differences*, *47*(8), 900–905. https://doi.org/10.1016/j.paid.2009.07.012

Embretson, S., & Yang, X. (2006). 23 Automatic item generation and cognitive psychology. In *Handbook of statistics* (Vol. 26, pp. 747–768). Elsevier. https://doi.org/10.1016/S0169-7161(06)26023-1

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. Psychometrika, *64*(4), 407–433. https://doi.org/10.1007/BF02294564

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233. https://doi.org/10.1037/h0057532

Gierl, M. J., & Lai, H. (2013). Instructional Topics in educational measurement (ITEMS) module: Using automated processes to generate test items. *Educational Measurement: Issues and Practice*, *32*(3), 36–50. https://doi.org/10.1111/emip.12018

Goldberg, L. R. (Oregon Research Institute). (2018). (2,8,10 & others) International Personality Item Pool (IPIP) [Data set]. Harvard Dataverse. https://doi.org/10.7910/DVN/UF52WY

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

Hinkin, T. R. (1995). A Review of Scale Development Practices in the Study of Organizations. Journal of Management, 21(5), 967–988. https://doi.org/10.1177/014920639502100509

Johnson, J. A. (2001). Personality psychology: Methods. In N. J. Smelser, & P. B. Baltes (Eds.), *International encyclopedia of the social & behavioral sciences* (pp. 11313–11317). Pergamon. https://doi.org/10.1016/B0-08-043076-7/01671-5

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. Journal of Research in Personality, 39(1), 103–129. https://doi.org/10.1016/j.jrp.2004.09.009

Kashdan, T. B., Stiksma, M. C., Disabato, D. J., Mcknight, P. E., Bekier, J., Kaji, J., & Lazarus, R. (2018). The five-dimensional curiosity scale: Capturing the bandwidth of curiosity and identifying four unique subgroups of curious people. *Journal of Research in Personality*, *73*, 130–149. https://doi.org/10.1016/j.jrp.2017.11.011

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, *9*(2), 202–220. https://doi.org/10.1177/1094428105284919

Larochelle, H., Erhan, D., & Bengio, Y. (2008). Zero-Data Learning of New Tasks. Proceedings of the 23rd National Conference on Artificial Intelligence, 2, 646–651.

Mastrich, Z., & Hernandez, I. (2021). Results everyone can understand: A review of common language effect size indicators to bridge the research-practice gap. *Health Psychology*, *40*(10), 727–736. https://doi.org/10.1037/hea0001112

Morris, J., Lifland, E., Yoo, J. Y., Grigsby, J., Jin, D., & Qi, Y. (2020). TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 119–126. https://doi.org/10.18653/v1/2020.emnlp-demos.16

Ni, J., Ábrego, G. H., Constant, N., Ma, J., Hall, K. B., Cer, D., & Yang, Y. (2021). Sentence-T5: Scalable sentence encoders from pre-trained text-to-text models. *Advances in Neural Information Processing Systems: ACL 2022*, 1864–1874 https://arxiv.org/abs/2108.08877

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 3982–3992. https://doi.org/10.18653/v1/d19-1410

Reimers, N., & Gurevych, I. (2020). Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 4512–4525. https://doi.org/10.18653/v1/2020.emnlp-main.365

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. ArXiv:1910.01108 [Cs]. http://arxiv.org/abs/1910.01108

Sherer, M., Maddux, J. E., Mercandante, B., Prentice-Dunn, S., Jacobs, B., & Rogers, R. W. (1982). The Self-efficacy scale: Construction and validation. *Psychological Reports*, *51*(2), 663–671. https://doi.org/10.2466/pr0.1982.51.2.663

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*(1), 117–143. https://doi.org/10.1037/pspp0000096

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*(2), 173–180. https://doi.org/10.1207/s15327906mbr2502_4

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, 6000–6010.

Von Davier, M. (2018). Automated item generation with recurrent neural networks. *Psychometrika*, *83*(4), 847–857. https://doi.org/10.1007/s11336-018-9608-y

Zoph, B., Yuret, D., May, J., & Knight, K. (2016). Transfer learning for low-resource neural machine translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 1568–1575. https://doi.org/10.18653/v1/D16-1163

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Hernandez, I., & Nie, W. (2022). The AI-IP: Minimizing the guesswork of personality scale item development through artificial intelligence. *Personnel Psychology*, 1–24. Advance online publication. https://doi.org/10.1111/peps.12543

APPENDIX A: Comparison of original seed items from Big Five Inventory 2 Scale and AI-generated items

Facet	Original seed items	AI-generated items (Sample)
Agreeableness	• Is compassionate has a soft heart • Is respectful, treat others with respect • Has a forgiving nature	• Am kind-hearted • Am a warm person • Feel little compassion for other people
Conscientiousness	• Is systematic likes to keep things in order • Is persistent works until the task is finished • Is reliable can always be counted on	• Do everything neat and organized • Can always be depended on • Will often leave a mess
Extraversion	• Is outgoing sociable • Is dominant acts as a leader • Is full of energy	• Am a very energetic and enthusiastic person • Have strong leadership • A a shy, introverted person
Neuroticism	• Worries a lot • Often feels sad • Is temperamental and gets emotional easily	• Often find myself upset • Often have intense mood swings • Rarely am bothered by things
Openness to experience	• Is complex a deep thinker • Is fascinated by art music or literature • Is original comes up with new ideas	• Have an eclectic imagination • Think very creatively • Have little interest in intellectual or creative pursuits

Note: The sample AI-generated items include only a subset of the final items used.

APPENDIX B: Comparison of original seed items from Self-Efficacy Scale and AI-generated items

Facet	Original seed items	AI-generated items (Sample)
General self-efficacy	• If I can't do a job the first time, I keep trying until I can • When I make plan, I am certain I can make them work • Failure just makes me try harder	• Keep at it until I succeed • Am sure that my plans will succeed • Have little confidence in my ability to get things done
Social self-efficacy	• When I'm trying to become friends with someone who seems uninterested at first, I don't give up easily • I have acquired my friends through my personal abilities at making friends • If I see someone I would like to meet, I go to that person instead of waiting for him or her to come to me	• Am good at maintaining friendships • Love to get to know someone • Find it difficult to initiate new friendships

Note: The sample AI-generated items include only a subset of the final items used.

APPENDIX C: Comparison of original seed items from Curiosity Scale and AI-generated items

Facet	Original seed items	AI-generated items (Sample)
Joyous Exploration	• I seek out situations where is likely that I will have to think in depth about something • I enjoy learning about subjects that are unfamiliar to me • I find it fascinating to learn new information	• Consider myself an intellectually curious person • Love to discover things • Have little curiosity about many topics
Deprivation Sensitivity	• I can spend hours on a single problem because I just can't rest without knowing the answer • I feel frustrated if I can't figure out the solution to a problem, so I work even harder to solve it • I work relentlessly at problems that I feel must be solved	• I try very hard to do things that others find difficult • Am a "problem solver" • Find it difficult to get myself to do things.
Stress Tolerance	• The smallest doubt can stop me from seeking out new experiences • I cannot handle the stress that comes from entering uncertain situations • I find it hard to explore new places when I lack confidence in m abilities	• Can find it difficult to relax • Fear not being able to solve a difficult problem • Am usually able to think of a way to solve my problems
Social Curiosity	• I like to learn about the habits of others • When other people are having a conversation, I live to find out what it's about • When around others, I like listening to their conversations.	• Love to hear about others' experiences • Am fascinated by the psychology of others • Dislike to talk about personal things
Thrill Seeking	• Risk-taking is exciting to me • When I have free time, I want to do things that are a little scary • Creating and adventure as I go is much more appealing than a planned adventure	• Love the thrill of the unexpected • Enjoy a "real-life adventure" • Prefer a dull life

Note: The sample AI-generated items include only a subset of the final items used.