ABSTRACT

PAUL, STEPHEN CHRISTIAN. Prompt Engineering in Scale Development: Leveraging Large Language Models for I-O Psychology Research (Under the direction of Dr. Adam Meade).

Large language models (LLMs) have emerged as powerful tools in applied psychology, leading to growing interest among researchers and practitioners in how to use them effectively. Scale development represents a key area of LLM application, and understanding their impact on this foundational process remains an important research priority. This study examined the use of LLMs for automatic item generation (AIG) in the development of job attitude scales. Specifically, the study compared three distinct prompting approaches for AIG and evaluated the resulting scales based on their psychometric properties. Previous research on AIG has primarily relied on fine-tuning LLMs using large item datasets, has not compared the effectiveness of different prompting techniques, and has focused almost exclusively on generating personality items. The current study addresses these gaps by using zero-shot prompting, few-shot prompting, and explanation-based prompting to generate items for three constructs: job satisfaction, organizational commitment, and organizational justice. Each resulting scale was assessed for reliability, content validity, construct validity, and criterion-related validity. In Phase 1, 90 naïve judges rated how well each item corresponded to its construct definition to evaluate content validity. Explanation-based prompts produced scales with the highest ratings of definitional correspondence, followed by zero-shot prompts, while few-shot prompts performed slightly worse. In Phase 2, construct validity, criterion-related validity, and reliability were assessed using a final sample of 346 participants recruited from Prolific. Participants were randomly split into two sub-samples: one for exploratory factor analysis (EFA) and one for confirmatory factor analysis (CFA). Across prompting techniques, GPT-authored items generally aligned with their

intended factor structures. Among scales generated by the three prompting techniques, explanation-based scales resulted in the best-fitting models, zero-shot scales performed reasonably well, and few-shot scales showed comparatively weaker fit. In terms of nomological network validity, GPT-authored scales (zero-shot, few-shot, explanation-based) of job satisfaction, organizational commitment, and organizational justice functioned as hypothesized within established theoretical frameworks. Across the overall sample, the automatically generated scales exhibited strong reliability.

© Copyright 2025 by Stephen C. Paul

All Rights Reserved

Prompt Engineering in Scale Development: Leveraging Large Language Models for I-O Psychology Research

by
Stephen Christian Paul

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Psychology

Raleigh, North Carolina

2025

APPROVED BY:

_____	_____
Dr. Adam Meade Committee Chair	Dr. Lori Foster
_____	_____
Dr. Mark Wilson	Dr. S. Bartholomew Craig

BIOGRAPHY

Stephen Paul earned a Bachelor of Science in Economics from Texas Christian University before completing a Ph.D. in Industrial-Organizational Psychology at North Carolina State University. While at NC State, he conducted research in the 4D Lab under the mentorship of Dr. Lori Foster, focusing on the intersection of industrial-organizational psychology and emerging technologies. Additionally, he served as a graduate research assistant at The Laboratory for Analytic Sciences (LAS) under the supervision of Dr. Mark Wilson, where he collaborated on the development of a data-driven tool designed to enhance intelligence analysis by capturing team data, identifying misalignments, and providing insights into performance-related challenges. He currently works as a Consultant at PDRI by Pearson where he develops and validates selection assessments.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. Lori Foster, for her consistent mentorship and guidance throughout my graduate studies. Dr. Foster's wisdom, kindness, and dedication to her students is exceptional. She met with me weekly, offered detailed and insightful feedback on my work, and consistently helped me make progress toward my goals. I feel incredibly fortunate to have had her as my advisor during my time at NC State.

I am also deeply grateful to Dr. Mark Wilson for the opportunity to work with him at the Laboratory for Analytic Sciences. His expertise, intellectual curiosity, and sense of humor made him both an inspiring mentor and a joy to collaborate with. I would also like to thank my dissertation chair, Dr. Adam Meade, for his supervision during the dissertation writing process. His ability to make complex topics accessible was both impressive and deeply appreciated.

I would also like to thank my cohort—Jason Saville, Madiha Qasim, Liz Tracy, and Sara Stegemoller—for their friendship and support. A special thanks to the students from senior cohorts who helped guide me through my first years of graduate school: Jenna McChesney, Demetrius Green, and English Sall.

Finally, I would like to thank my family—Sarah, for always instilling confidence in me; Hannah, for offering many wise insights and perspective along the way; and my mother, for her steadfast love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The development of artificial intelligence (AI) has and will continue to have a significant impact on industrial-organizational (I-O) psychology. One affected area is traditional item generation due to the increased study and application of automatic item generation (AIG). AIG is the process by which AI generates items through computer technology (Gierl et al., 2021). There are two kinds of automatic item generation: template-based AIG and non-template-based AIG (Shin, 2021). Template-based AIG relies on templates generated by subject matter experts (SMEs) to direct the creation of items (Gierl et al., 2021). Non-template-based AIG relies on natural language processing (NLP) methods to generate content (Baghaee, 2017; Yao & Zhang, 2010). In NLP, large language models (LLMs) have seen important developments, with LLMs such as ChatGPT gaining 100 million users in its first two months (Hu, 2023). In I-O research, large language models (LLMs) have already been used for scale development purposes, such as item generation (e.g., Lee et al., 2023) and content analysis (Fyfee et al., 2023).

One milestone in the development and widespread use of LLMs is the increased usability enabled by reinforcement learning from human feedback (RLHF). RLHF was used to fine-tune GPT-3 (generative pre-trained transformer) to follow written human instructions (Ouyang et al., 2022) leading to a class of "InstructGPT' models that allow users to effectively communicate task instructions through natural language, rather than programming language. Before the release of GPT-3, the conventional method for using language models entailed fine-tuning the model on a subset of a task dataset. However, GPT-3 changed traditional methods by attaining strong performance across multiple tasks without the use of fine-tuning. Instead, GPT-3 utilized a few-shot prompt approach, where a small number of examples of completed tasks were provided as part of the input to the trained model (Brown et al., 2020).

This led to a new area of research on human communication with large language models called prompt engineering (Efrat and Levy, 2020). Reynolds and McDonell (2021) define prompt engineering as "programming in natural language [in which] there are boundless number of functions that we know intimately but do not have names for." Prompt engineering refers to the design and creation of input prompts for the model to generate text that aligns with an intended task. Despite significant promise in this area, the existing methods for directing models to perform new tasks are constrained (Min et al., 2022). Recent studies have explored few-shot (i.e., prompt with a few examples) and zero-shot (i.e., prompts without examples) approaches as alternatives to fine-tuning models showing success for various tasks (e.g., Reynolds & McDonell, 2021; Min et al., 2021). However, no studies have compared or systematically evaluated these different approaches for the task of item generation. The goal of this study is to explore how different prompting techniques influence the automatic generation of items and to determine which approaches are best suited for scale development.

AIG using Large Language Models

Previous studies exploring the use of LLMs for scale development have primarily relied on large pre-existing datasets of items to train the models. They also focus almost exclusively on personality constructs (e.g., The Big Five) (von Davier, 2018; Hommel et al., 2021; Hernandez & Nie, 2022; Lee et al., 2023).

Von Davier (2018) was the first to successfully use non-template-based AIG to generate personality items using a neural network-based machine learning approach. He trained a language model on validated personality items from the International Personality Item Pool (IPIP; Goldberg, 1999), selected and combined 24 of the automatically generated items with 17 published items from the IPIP, and presented these items randomly to participants on Amazon

Mechanical Turk. Items loaded according to the intended construct (e.g., openness to experience) more often than to the method of their generation (e.g., automatically or human-authored).

Hommel et al. (2021) followed up on this study and the use of non-template-based automatic item generation. Improving upon the previous approach, they took advantage of new developments in language models (i.e., using a transformer-based deep learning model) when producing items for personality traits. They had three sets of items 1) items introduced during fine-tuning, trained with five constructs (openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism) 2) a second set of items (i.e., benevolence, egalitarianism, and pessimism) that were not included in fine-tuning, and 3) human-authored items from a previously validated IPIP scale (Goldberg, 1999; Hommel et al., 2021). Overall, they found evidence that most automatically generated items exhibited similar psychometric properties to human-authored items, while some exhibited poorer factor loadings and reliability (Hommel et al., 2021).

Hernandez and Nie (2022) used GPT-2 to identify the semantic structures of an item bank of IPIP items (Goldberg, 1999), then created 1,000,000 new items. After generating these items, they examined item quality for non-redundancy, range of words, and congruence in structure with the human-authored items, detecting no major differences between human and AI -authored items. They also compared psychometric properties across human and machine generated items, finding the AI-generated scales correlated, on average, around .72 with the human-generated scales. However, the correlations ranged from .39 to .72 on the sub-facets of scales that did not have equivalent constructs to the IPIP.

Lee et al. (2023) used few-shot prompting to generate items based on example items for their respective constructs. The researchers also took a more rigorous approach than previous

studies in analyzing the psychometric properties of the items by examining content validity, construct validity, criterion validity, and differential item functioning. The machine-authored personality test was found to effectively measure the Big Five personality traits and had comparable reliability to the human-authored test. Additionally, it displayed similar convergent and discriminant validity and had similar criterion-related validity to the human-authored items. They used an early version of GPT-3 (Brown et al., 2020), which preceded the "InstructGPT" models (e.g., text-davinci-002) that were fine-tuned to follow human instructions using reinforcement learning with human feedback (Ouyang et al., 2022). As a result, the model was less responsive to natural language instructions and less aligned with human intent. This may not have posed a significant limitation for few-shot prompts, where examples provide structure and context. However, it could have limited performance in prompting techniques that rely more heavily on natural communication, such as explanation-based prompts, chain-of-thought reasoning, or detailed zero-shot instructions. Lee et al.'s (2023) few-shot prompt included five traits, each accompanied by 25 example items. Their approach requires access to well-developed item banks, which may limit its applicability in contexts where such resources do not yet exist.

Prompts

Few-Shot and Zero-Shot Prompts

LLMs can perform many different tasks by either providing the model with examples (known as few-shot prompt) or only giving the model task instructions (referred to as a zero-shot prompt) (Kojima et al., 2022). Giving an input to the model is called a prompt. These prompts can be manually created and tailored to specific tasks, as shown by Schick and Schütze (2021). Prompt engineering refers to the process of designing and creating these prompts for improved performance.

Few-shot prompting, also referred to as “task specification by demonstration,” is more suited to tasks that require a very specific format, where context is lacking, or when high-quality examples are available (Reynolds & McDonell, 2021, p. 5). Multiple studies have shown that few-shot prompting outperforms zero-shot prompting on various tasks (Brown et al., 2020; Zhao et al., 2021). Zero-shot learning has also been tested in pre-trained language models and has successfully performed many tasks (Brown et al., 2020; Wei et al., 2022). A zero-shot prompt assumes that the language model has already learned the necessary functions to carry out the task. With zero-shot prompts, the language model uses knowledge of language patterns to generate new items that are likely to be aligned with the intended construct.

Reynolds and McDonnell (2021) made the case that few-shot prompts could lead to worse performance in certain cases because the primary function of examples is that of task location, not in-context learning. Therefore, it could be argued superior task location is better achieved through a well-engineered prompt rather than labeled data (Reynolds & McDonell, 2021). Further research is needed on prompt engineering to understand why certain prompts are more effective than others (Min et al., 2022).

Explanation-Based Prompts and Task Instructions

There are other methods of prompt engineering that extend beyond the decision of number of examples (e.g., zero-shot versus few-shot). Other techniques such as chain-of-thought prompts (Wei et al., 2022) and in-context explanations (Lampinen et al., 2022) have been explored for the purpose of enhancing model performance. Lampinen et al. (2022) demonstrated that the addition of in-context explanations can considerably improve the capability of the models. They showed that a few-shot prompt with examples and explanations led to superior performance over just using examples. Mitchell (2021) supported this notion asserting that

LLMs can benefit from explanations. Although explanations show early success, researchers have yet to determine which types of explanations are most effective.

Another way of conceptualizing explanations in prompt design can be centered around the idea of "constraining behavior," as introduced by Reynolds and McDonell (2021). Reynolds and McDonell (2021) encapsulate this approach with the question: "What prompt will result in the intended behavior and *only* the intended behavior?" (p. 6). In this approach, prompts should guide model output by limiting interpretive ambiguity and directing the model toward specific outcomes. Because language models can pursue many plausible continuations, variations in prompt wording can lead to significantly different outputs. From this perspective, explanations act as constraints that help narrow the model's direction of reasoning and generation. Reynolds and McDonell (2021) found that carefully written task instructions outperformed few-shot examples in some cases because examples can introduce "semantic contamination" (p. 3), where the model interprets the examples as part of a narrative rather than as independent demonstrations of a function. In such cases, the model may infer continuity between examples, leading it to generate responses based on imagined connections rather than the intended task structure. Clear task instructions and explanations, if done well, directly communicate the desired behavior and reduce ambiguity, helping the model focus on the task without being influenced by unintended patterns. This is consistent with findings from Liu et al. (2021), who showed that explicit task instructions and descriptions improve LLM performance by providing clearer signals about the intended behavior.

Hypothesis Development

AIG of Non-Personality I-O Constructs

As previously noted, studies have almost exclusively largely focused on personality constructs with large existing datasets (von Davier, 2018; Hommel et al., 2021; Hernandez & Nie, 2022; Lee et al., 2023). Based on the current non-template based AIG literature, we know a lot about how language models generate items related to personality, but little beyond this. Model performance may be overestimated if similar items were encountered during pre-training, a concern heightened by the widespread availability of Big Five personality items in public datasets. If few-shot learning is primarily effective because of task location, these models have an advantage on constructs that have substantial representation in their pre-training corpus.

This study explores three constructs that are highly relevant to the field of industrial-organizational psychology: job satisfaction, organizational commitment, and organizational justice. These constructs were selected based on their centrality in the nomological network of industrial-organizational psychology, their well-defined nature (making them suitable for few-shot and explanation-based prompts), their diverse factor structures, and their distinction from personality constructs.

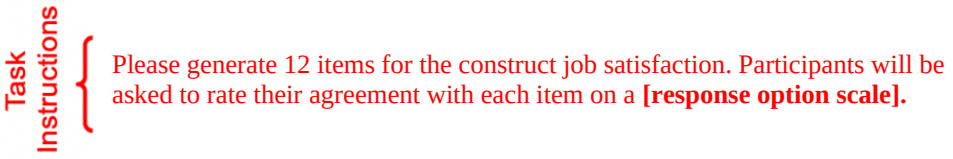
The availability of items and definitions is relevant because this information was directly used in the prompts. In the few-shot prompt, items from pre-existing scales served as examples, while in explanation-based prompts, construct definitions were incorporated as a form of explanation.

Zero-Shot Prompts for I-O Job Attitudes. No zero-shot prompts have been tested in scale development. Zero-shot prompts simply require that we communicate instructions to the model; however, they do not specify how to perform the task or its purpose. Pre-trained LLMs

have been trained on large amounts of text data, including journal articles, books, and online discussions. To generate items for a construct, GPT-4 leverages its learned representations of language patterns to generate new sentences that are similar in structure and meaning to the input prompt. It could also use its knowledge of common phrasing and vocabulary used in the domain of interest to generate items that are relevant and appropriate.

Job satisfaction, organizational commitment, and organizational justice are each present in the vast amounts of text data, from its many conceptualizations and definitions in journal articles to its expressions in online discussions. In a zero-shot prompt, the LLM receives minimal guidance on what to focus on, relying on its pre-trained knowledge and learned representations to generate a response. See Figure 1 below for an example:

Figure 1
Zero-Shot prompt with only task instructions.



Note. Zero-shot prompting (only task instructions)

The choice of items generated by the LLM based solely on the construct label remains an open question. Given the lack of explicit guidance in a zero-shot prompt and the potential variability in how the model interprets the construct, I do not propose any specific hypotheses regarding the factor structure. Therefore, I propose the following research question:

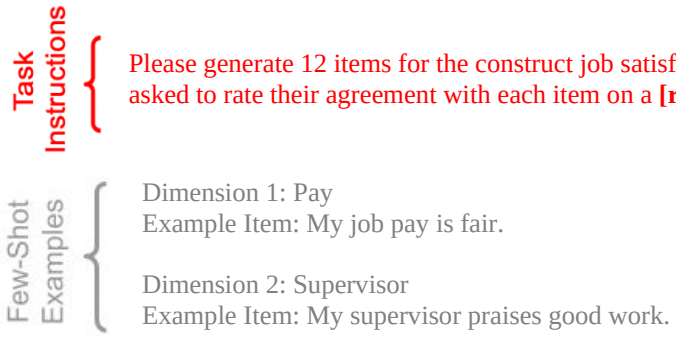
> **RQ-1:** What factor structure emerges for [job satisfaction, organizational commitment, organizational justice] items generated with a zero-shot prompt?

Few-Shot Prompts for Job Attitudes. One study has used a few-shot prompt approach for scale development (Lee et al. 2023). In the context of generating items for a construct using

few-shot examples, the repetition of the same item structure with different parameters is a common pattern in natural language and allows the language model to generate a variety of items for the same construct in a more efficient way. Lee et al. (2023) used 25 items as examples from a large pre-existing set of items (IPIP; Goldberg, 1999).

For the few-shot prompts, I drew example items from well-researched and commonly used scales in the I-O literature. I use the abridged Job Descriptive Index (Stanton et al., 2002), Meyer et al.'s (1993) organizational commitment scale, and Colquitt et al.'s (2001) organizational justice scale. These scales contain multiple dimensions, for example, Colquitt et al. (2001) outlines 4 dimensions of justice: procedural justice, distributive justice, interpersonal justice, and informational justice. In this case, the few-shot prompt specifies the dimension and its associated example item. See Figure 2 for an example prompt:

Figure 2
Few-shot prompt with task instructions and examples.



Note. Few-shot prompting (task instructions + example items)

As the prompt is structured according to the theoretical factor structure of the construct, I hypothesize that GPT-authored items will follow the provided factor structure Therefore, I propose the following hypotheses:

> **Hypothesis 1a**: A five-factor model structure for [few-shot, explanation-based] GPT-authored job satisfaction will fit significantly better than a one-factor solution.

Hypothesis 1b: A three-factor model structure for [few-shot, explanation-based] GPT-authored organizational commitment will fit significantly better than a one-factor solution.

Hypothesis 1c: A four-factor model structure for [few-shot, explanation-based] GPT-authored organizational justice will fit significantly better than a one-factor solution.

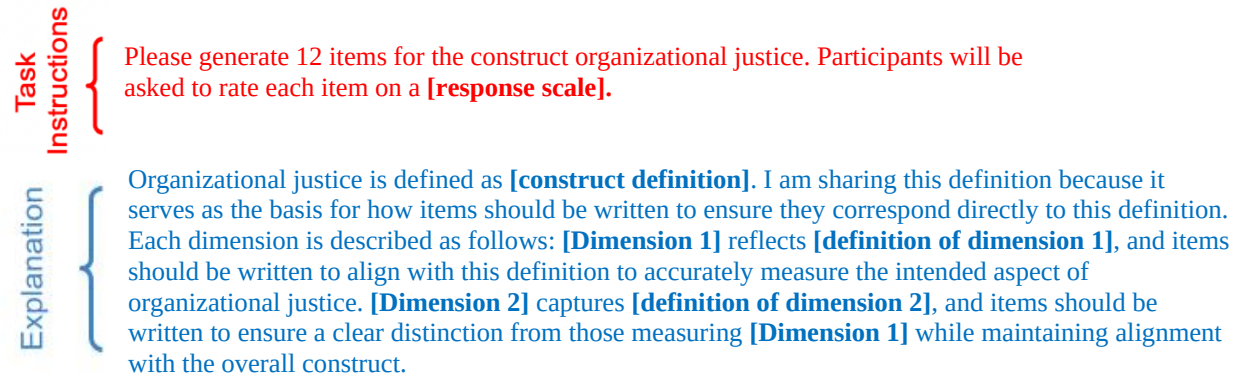
Explanation-Based Prompts for Job Attitudes. No studies have examined the effectiveness of explanations within prompts for AIG. Lee et al. (2023) introduced this as a future direction and provided one example item generated from an in-context explanation, incorporating context (e.g., sports) along with more detailed instructions to create an item reflecting openness in sports situations.

In the present study, I drew upon the framework of prompting as a means of constraining behavior (Reynolds & McDonell, 2021) to guide the design of "explanation-based" prompts. Within the context of scale development, prompts can initiate content generation through task instructions and attempt to further shape the semantic boundaries of the model's outputs through explanations. Definitions of the construct and its dimensions can function as behavioral constraints, narrowing the range of plausible responses the model considers appropriate. Constructs like job satisfaction often appear in a wide range of conceptual forms across pretraining data, including academic literature, organizational reports, and online discourse. If a definition is not provided, the model may generate content based on inconsistent understandings of the construct. As part of the prompt, I provided an explanation to the model about why the definition was included and how it should be used. The prompt stated that the definition was intended to guide item generation and instructed the model to produce content that directly aligned with it. This was done to further constrain the model's potential behavior and ensure that the construct was interpreted and applied consistently and as intended.

I incorporated construct and dimension definitions from the abridged Job Descriptive Index (Stanton et al. 2002) for job satisfaction, Meyer et al.'s (1993) scale for organizational commitment, and Colquitt et al.'s (2001) scale for organizational justice. These definitions were used as explanations within the prompts, with the goal of helping the model differentiate constructs and dimensions and classify items accurately. Figure 3 provides an example prompt that incorporates definitions along with additional contextual information that can be included in an explanation-based prompt to enhance construct clarity.

Figure 3

Explanation-based prompt with task instruction and explanations.



Note. Explanation-based prompting (task instructions + explanation).

By explicitly defining the construct, its dimensions, and the factor structure, this prompt template provides a structured explanation to guide item generation. The explanation provides definitions and clarifies their purpose by ensuring that the model generates items that correspond directly to the construct. It also specifies that items should align with their respective dimensions while remaining distinct from one another. The goal is to establish clear guidelines so that generated items accurately reflect the intended construct and maintain conceptual distinctions between dimensions.

Explanation-based prompting for job satisfaction used the definition (e.g., degree to which the employee is satisfied with how their supervisor supervises and/or the qualities the supervisor possesses) and factor structure (work, pay, promotion, supervisor, and co-workers) from the abridged JDI (Stanton et al., 2002). I hypothesize that the items generated will follow a five-factor structure that corresponds with the five JDI dimensions. Therefore, I propose the following hypothesis:

> **Hypothesis 2a**: A five-factor model structure for [few-shot, explanation-based] GPT-authored job satisfaction will fit significantly better than a one-factor solution.

Explanation-based prompting for organizational commitment used the definition by Meyer and Allen (1991), which refers to organizational commitment as an individual's psychological connection to their organization. Meyer and Allen (1991) proposed a three-part framework for understanding organizational commitment, which comprises affective commitment, continuance commitment, and normative commitment. Affective commitment relates to an individual's emotional investment in the organization and its objectives, continuance commitment is based on a person's belief that staying with the organization is necessary for economic or practical reasons, and normative commitment is rooted in an individual's ethical or moral obligation to remain with the organization (Meyer & Allen, 1991). As it relates to factor structure, I hypothesize that the items generated will follow a three-factor structure that corresponds with the three organizational commitment dimensions: affective commitment, continuance commitment, and normative commitment. Therefore, I propose the following hypothesis:

> **Hypothesis 2b**: A three-factor model structure for [few-shot, explanation-based] GPT-authored organizational commitment will fit significantly better than a one-factor solution.

Explanation-based prompting for organizational justice used Colquitt et al. (2001) four component model of justice: distributive justice, procedural justice, interactional justice, and informational justice. As it relates to factor structure, I hypothesize that the items generated will follow a four-factor structure that corresponds with the four organizational justice dimensions: distributive justice, procedural justice, interactional justice, and informational justice. Therefore, I propose the following hypothesis:

> **Hypothesis 2c**: A four-factor model structure for [few-shot, explanation-based] GPT-authored organizational justice will fit significantly better than a one-factor solution.

Nomological Network Validity

An important aspect of evaluating the quality of items generated by GPT is assessing construct validity by examining their correlations with related constructs and criterion-related validity by analyzing their relationships with relevant outcomes.

Job Satisfaction

As summarized by Judge et al. (2020), job satisfaction had a positive relationship with positive affect (PA) (Watson & Slack, 1993), and meta-analytic evidence also supported a moderate correlation between PA and job satisfaction, as well as between negative affect (NA) and job satisfaction (Thoresen et al., 2003). Additionally, meta-analytic findings indicated that neuroticism, extraversion, and conscientiousness were moderately correlated with job satisfaction (Judge, Heller, & Mount, 2002). Therefore, I propose the following hypotheses:

> **Hypothesis 3a**: GPT-authored job satisfaction will positively relate to positive affect, extraversion, and conscientiousness.
>
> **Hypothesis 3b**: GPT-authored job satisfaction will negatively relate to negative affect and neuroticism.

Criterion-Related Validity. Judge et al. (2020) detailed criterion-related validity evidence for job satisfaction across multiple meta-analyses. For example, there is support that

job satisfaction predicted organizational citizenship behavior (OCB) (Kinicki et al., 2002) and additional meta-analytic evidence that showed a relationship between aspects of job satisfaction, such as pay, coworkers, work, and supervision, and OCB (Kinicki et al., 2002). According to Dalal (2005), job satisfaction had a negative relationship with counterproductive work behavior (CWB). Therefore, I propose the following hypotheses:

> **Hypothesis 3c**: GPT-authored job satisfaction will positively relate to OCB.
> **Hypothesis 3d**: GPT-authored job satisfaction will negatively relate to CWB.

Although I expect a positive relationship between the GPT-authored and human-authored scales, the strength of relationship or presence of redundancy is unclear. As such, I ask the following research question to evaluate redundancy:

> **RQ-2***:* What is the correlation between GPT-authored and human-authored job satisfaction?

Organizational Commitment

Multiple studies have demonstrated a positive relationship between the personality traits of conscientiousness and extraversion and organizational commitment (Choi et al., 2015; Erdheim et al., 2006; Hackney, 2012). Locus of control has also shown a positive association with organizational commitment (Johnson, Chang, & Yang, 2010). Additionally, Meyer and Maltin (2010) and Klein and Park (2020) summarized evidence that showed commitment was positively associated with well-being and negatively related to outcomes of burnout and stress. Therefore, I propose the following hypotheses:

> **Hypothesis 4a***:* GPT-authored organizational commitment will positively relate to locus of control, conscientiousness, and extraversion.
>
> **Hypothesis 4b***:* GPT-authored organizational commitment will negatively relate to burnout and stress.

To evaluate redundancy, I ask the following research question:

RQ-3*:* What is the correlation between GPT-authored and human-authored organizational commitment?

Organizational Justice

Martínez-Tur et al. (2020) summarized construct validity evidence for organizational justice, noting a meta-analysis conducted by Ng & Sorensen (2009) that showed a link between positive and negative affect and organizational justice. Additionally, Shi et al. (2009) investigated the relationship between Big Five personality traits and organizational justice. Their study results showed a significant positive relationship with agreeableness and negative relationship with neuroticism. Therefore, I propose the following hypotheses:

Hypothesis 5a: GPT-authored organizational justice will positively relate to positive affect and agreeableness.

Hypothesis 5b: GPT-authored organizational justice will negatively relate to negative affect, neuroticism, stress, and burnout.

Criterion-Related Validity. Colquitt et al. (2013) found a relationship between organizational justice and task performance, OCB, and CWB. Therefore, I propose the following hypotheses:

Hypothesis 5c: GPT-authored organizational justice will positively relate to OCB.

Hypothesis 5d: GPT-authored organizational justice will negatively relate to CWB.

To evaluate redundancy, I ask the following research question:

RQ-4: What is the correlation between GPT-authored and human-authored organizational justice?

PHASE 1: AUTOMATIC ITEM GENERATION AND CONTENT VALIDITY

In phase 1, items were automatically generated using ChatGPT powered by GPT-4, and content validity was assessed by naïve judges.

Sample

To assess content validity, I recruited 90 participants from Prolific to act as naïve judges (Colquitt et al. 2019). I followed the same criteria employed by Colquitt et al. (2019) such that participants must be employed, live in the United States and be proficient in English. Additionally, participants were paid an hourly rate of $8.00 USD for their participation. Full instructions provided to participants can be found in Appendix A.

The average age of the participants was 35 years old, with ages ranging from 19 to 67. The gender distribution was as follows: 60% identified as women, 29% as men, 5% as gender queer, and 6% as other. The sample's racial composition included 69% White, 14% African American or Black, 10% Asian, 6% Two or More Races, and 1% American Indian or Alaska Native. The sample’s ethnic composition included 13% Hispanic or Latino/a and 87% Not Hispanic or Latino/a.

In terms of education, 25% of participants held a bachelor's degree, 22% completed some college, 22% held a master's degree, 11% had a high School diploma or equivalent, 10% had an associate’s degree, 2% had vocational Training, 2% had an applied or professional doctorate degree, 2% had some graduate school, and 1% had a doctorate degree. To assess data quality, I used one question adapted from Meade and Craig (2012) that read “Please consider how carefully you have responded to items in this study. In your honest opinion, should we use your data? You will be compensated regardless of your answer, but your honesty helps to improve our

work.” Response options were “No” (0) and “Yes” (1). Seven participants were removed for responding 'No,' resulting in a final sample of 83 participants.

Language Model

The Generative Pretrained Transformer (GPT)-3 language model was introduced in 2020 with 175 billion parameters (Brown et al., 2020). Its pre-training process incorporated multiple data sources, including WebText2, two internet-based book corpora (Books1 & Books2), Wikipedia, and additional undisclosed datasets. In the current study, I use ChatGPT, a successor to the “InstructGPT” models (Ouyang et al., 2022), which builds on OpenAI's GPT-3 family. Specifically, ChatGPT powered by GPT-4 was used, which incorporates fine-tuning through RLHF. However, the GPT-4 technical report does not disclose details on its training dataset composition, model architecture, hyperparameters, or other technical specifics, citing competitive considerations and safety concerns as the primary reasons for these omissions (OpenAI, 2023).

Automatic Item Generation

I used three different prompting techniques for automatic item generation with ChatGPT: a zero-shot prompt, a few-shot prompt, and an explanation-based prompt. The zero-shot prompt included task instructions directing the model to generate items for each construct and dimension without providing examples or additional explanation. Only the construct and dimension labels were provided. The few-shot prompt used the same task instructions but included three example items. I selected a three-shot format to reflect a realistic scenario in which only a small number of example items are available. The explanation-based prompt also included the task instructions, but instead of examples, it provided definitions of the construct and its dimensions to guide item generation. To ensure consistency, each prompting condition was conducted in a separate chat

session with model memory disabled. Table 1 presents the construct labels, example items, and definitions used across constructs and dimensions.

Content Validity

I gathered content validity evidence using Hinkin and Tracey's (1999) approach detailed within Colquitt et al. (2019). I calculated the average definitional correspondence or Hinkin Tracey correspondence (*htc*) by taking the judges rating of the degree to which scale items correspond to the construct's definition (1 = Item is an ***extremely bad*** match to the definition to 7 = Item is an ***extremely good*** match to the definition) divided by the number anchors *a* (Colquitt et al., 2019):

$$htc = \text{average definitional correspondence rating} / a \quad (1)$$

I used Colquitt et al. (2019) evaluation criteria for *htc* not normed to the average correlation between focal scale and orbiting scales. Orbiting constructs (i.e., similar but distinct theoretically and empirically) in this context are used to determine "the degree to which a scale's items correspond to the focal construct's definition, rather than the orbiting constructs definition" (Colquitt et al., 2019, p. 1243). According to Colquitt et al.'s (2019) guidelines for *htc* scores not normed to the average correlation between focal and orbiting scales, *htc* scores below .84 signal weak content adequacy, .84 to .86 have moderate content adequacy, and scores .87 and above signal strong content adequacy.

For GPT-authored job satisfaction, average item *htc*s ranged from .70 to .90 (see Table 2). The results showed that the zero-shot prompt resulted in an *htc* average of .83, the few-shot resulted in an average of .80, and the explanation-based prompt resulted in the highest *htc* average at .84 (see Table 5). For GPT-authored organizational commitment, average item *htc*s ranged from .57 to .89 (see Table 3). The overall *htc* performance across prompts was slightly

lower. Both the zero-shot and few-shot prompts produced an *htc* average of .76, while explanation-based prompting showed a slight improvement with an *htc* average of .77 (see Table 5). For GPT-authored organizational justice, average item *htc*s ranged from .64 to .87 (see Table 4). Similarly, the zero-shot prompt resulted in an *htc* average of .80, the few-shot prompt resulted in a lower *htc* average of .72, and the explanation-based prompt again had the highest *htc* average of .82 (see Table 5).

PHASE 2: FACTOR STRUCTURE AND NOMOLOGICAL NETWORK VALIDITY

In Phase 2, I assessed the construct validity, criterion-related validity, and reliability of GPT-authored scales of job satisfaction, organizational commitment, and organizational justice.

Sample

A total of 370 participants were recruited via Prolific and compensated at a rate of $8.00 per hour. I followed the same criteria employed in Phase 1 and specified by Colquitt et al (2019) such that participants must hold employment, currently live in the United States, and be proficient in English. The average age of the participants was 38 years old, with ages ranging from 20 to 69 years old. Gender distribution was as follows: 63% identified as women, 35% as men, and 2% as other. The sample's racial composition included 65% White, 21% African American or Black, 8% Two or more races, 4% Asian, and less than 1% for American Indian or Alaska Native, Middle Eastern or North African, Native Hawaiian or Other Pacific Islander, and Other. For ethnic composition, 8% identified as Hispanic or Latino/a and 92% selected not Hispanic or Latino/a. For education, 38% of participants held a bachelor's degree, 20% completed some college, 14% held a master's degree, 13% held a high school diploma, 8% held an associate's degree, 2% held an applied or professional doctorate degree, 2% had vocational Training, 2% had some graduate school, 1% had some high school, 1% had a doctorate degree, and less than 1% selected less than high school. See Appendix K for a full list of demographic questions.

A total of (n = 24) individuals were excluded from data collection and/or data analysis for failing one or two attention checks (n = 18) or answering "No" to a dichotomous item (n = 6) that asked participants whether their data should be used (Meade & Craig, 2012). This resulted in a final dataset of N = 346.

Unlike Cronbach's alpha, which assumes homogeneous item-construct relations, omega accounts for the possibility of heterogeneous item-construct relationships (Hayes & Coutts, 2020). The omega scores for the GPT-authored job satisfaction scales were 0.92, 0.92, and 0.93, for zero-shot, few-shot, and explanation-based, respectively. The omega scores for the GPT-authored organizational justice scales were 0.96, 0.95, and 0.96, for zero-shot, few-shot, and explanation-based, respectively. The omega scores for the GPT-authored organizational commitment scales were 0.94, 0.96, and 0.94, respectively. This indicates high internal consistency across each GPT-authored scale.

Sub-Sample Split

The sample was randomly split into two sub-samples. On the first sub-sample, I conducted EFAs to explore the factor structure of zero-shot GPT authored scales and on the second sub-sample, I performed CFAs to confirm the factor structure of all GPT-authored scales across prompting techniques. Of the 346 remaining cases, Hinkin (1998) recommends a minimum of 150 responses for an EFA and 200 for a CFA. 150 participants were randomly split into the EFA sub-sample, and the remaining 196 cases were randomly assigned to the CFA sub-sample.

Procedure

Participants signed up for the study via Prolific and responded to a listing to complete a study on job attitudes. Participants were then redirected to a Qualtrics survey and asked to complete an informed consent page. If they agreed to the conditions of the informed consent, participants then responded to a variety of scales about their job attitudes, personality, and demographics. The survey included two attention checks and one honesty question adapted from

Meade and Craig (2012). The study took approximately 21 minutes to complete. The section below explains what measures were included in the study.

Measures

Attention Checks

I used two instructed response items to assess attention. One attention check item mirrors the adjacent items from the IPIP scale, reading "please select somewhat agree for this item" with possible responses ranging from 1 = strongly disagree to 7 = strongly agree. The other attention check item mirrored the adjacent items from the job satisfaction scale, reading "please select no for this item" with possible responses of 1 = Yes, 2 = ? and 3 = No.

GPT-Authored Job Satisfaction

Three separate scales for GPT-authored job satisfaction were generated: zero-shot GPT-authored job satisfaction, few-shot GPT-authored job satisfaction, and explanation-based GPT-authored job satisfaction. I used a 3-point scale (Yes, ?, No) to measure job satisfaction, ensuring consistency with the abridged Job Descriptive Index (Stanton et al., 2002). A full list of items can be found in Table 2.

GPT-Authored Organizational Commitment

Three separate scales for GPT-authored organizational commitment were generated: zero-shot GPT-authored organizational commitment, few-shot GPT-authored organizational commitment, and explanation-based GPT-authored organizational commitment. I used a 7-point Likert-type scale (1 = strongly disagree to 7 = strongly agree) to measure each dimension. A full list of items can be found in Table 3.

GPT-Authored Organizational Justice

Three separate scales for GPT-authored organizational justice were generated: zero-shot GPT-authored organizational justice, few-shot GPT-authored organizational justice, and explanation-based GPT-authored organizational justice. I used a 5-point scale with (1 = to a small extent to 5 = to a large extent) to measure organizational justice across 4 dimensions (distributive justice, procedural justice, interactional justice, and informational justice). A full list of items can be found in Table 4.

Job Satisfaction

I used the abridged Job Descriptive Index with a 7-point Likert-type scale (1 = Yes, 2 = ? 3 = No) to measure job satisfaction across 5 dimensions (work, pay, promotion, supervision, and coworkers) (Stanton et al., 2002). A full list of items can be found in Appendix B.

Organizational Commitment

I used a 7-point Likert scale (1 = strongly disagree to 7 = strongly agree) to measure three dimensions of organizational commitment (affective commitment, normative commitment, and continuance commitment) (Meyer et al.1993). A full list of items can be found in Appendix C.

Organizational Justice

I used a 5-point scale with (1 = to a small extent to 5 = to a large extent) to measure 4 dimensions (procedural, distributive, interpersonal, informational) of justice (Colquitt, 2001). A full list of items can be found in Appendix D.

Negative and Positive Affect

I used a 5-point scale with (1 = strongly disagree to 5 = strongly agree) to measure positive and negative affect schedule (PANAS) (Watson et al., 1988). A full list of items can be found in Appendix E.

Personality

I used a 20-item scale to measure extraversion, conscientiousness, neuroticism, openness to experience, and agreeableness using a 7-point Likert-type scale (1 = strongly disagree, 7 = strongly agree) (Mini-IPIP; Donnellan et al., 2006). A full list of items can be found in Appendix F.

Burnout

I used a three-item scale to measure burnout (Boswell et al., 2004). A sample item is "Today, I felt burned out from my work." A full list of items can be found in Appendix G.

Job Stress

I used a two-item measure of job stress from Solomon et al. (2022). The items include: "My job is more stressful than I had ever imagined" and "I fear that the amount of stress in my job will make me physically ill." A full list of items is provided in Appendix H.

OCB and CWB

I measured OCB and CWB targeted towards a respondent's organization, coworkers, and immediate supervisors with 6 items for each dimension (Dalal et al., 2009). An example item for OCB is "I spoke highly about my organization to others". An example item for CWB is "I behaved in an unpleasant manner towards my supervisor. A full list of items can be found in Appendix I.

Locus of Control

I used a 6-item scale to measure locus of control (Goldberg, 1999). An example item is "I believe that unfortunate events occur because of bad luck." A full list of items can be found in Appendix J.

Results

Sub-Sample 1

Exploratory factor analyses were conducted using the lavaan (0.5–20) package (Rosseel, 2012) in R (version 4.2.0). Maximum likelihood extraction was used, with Promax rotation (a method of oblique rotation). To determine the optimal number of factors to extract for each zero-shot GPT-authored scale, I used multiple criteria: inspection of scree plots, parallel analysis (Horn, 1965), the percentage of variance explained, and the principle of simple structure.

For the zero-shot job satisfaction scale, parallel analysis indicated a five-factor solution. An EFA confirmed this structure, with items loading onto their intended factors and displaying simple structure (e.g., supervision items loaded together on a single factor with minimal cross-loadings; see Table 6). For zero-shot organizational commitment scale, parallel analysis suggested a three-factor solution. An EFA supported this structure, with items again loading onto their intended factors and demonstrating simple structure (see Table 7). For the zero-shot organizational justice scale, items were generated to reflect four dimensions (distributive, procedural, informational, and interactional). However, parallel analysis indicated a three-factor solution. To further investigate, an EFA was conducted. Procedural justice items did not load onto a distinct factor but instead distributed across the distributive and informational justice factors (see Table 8).

Sub-Sample 2

On the second sub-sample, CFAs were conducted to evaluate the hypothesized factor structures of the zero-shot, few-shot, and explanation-based GPT-authored scales. For each scale, three models were tested: (1) the hypothesized factor model (e.g., a five-factor model for job satisfaction), (2) a one-factor model, and (3) a hierarchical model (e.g., five first-order factors

and one second-order factor for job satisfaction). Model fit was assessed using multiple indices, including chi-square (χ^2), comparative fit index (CFI), Tucker–Lewis index (TLI), root-mean-square error of approximation (RMSEA), and standardized root-mean-square residual (SRMR). Criteria for good fit were defined as CFI and TLI ≥ .95, RMSEA ≤ .06, and SRMR ≤ .08 (Hu & Bentler, 1999). While these cutoff values offer useful guidelines, they should not be applied rigidly. Fit indices should be interpreted in combination, as each provides unique information about model performance to allow for a more holistic evaluation of fit, considering sample size, model complexity, and theoretical considerations. I also conducted Likelihood Ratio Tests (LRTs) to statistically test differences in model fit between nested models, such as comparing a five-factor model to a one-factor model.

Job Satisfaction Factor Structure. Table 9 displays model fit indices for CFAs of GPT-authored job satisfaction scales developed using zero-shot, few-shot, and explanation-based prompting. Each scale was tested with one-factor, five-factor, and hierarchical structures. Based on theoretical expectations, I hypothesized that a five-factor model would provide better fit than a one-factor solution for the few-shot and explanation-based scales.

For the few-shot scale, the five-factor model showed mixed evidence of fit. Some indices approached recommended thresholds, including a CFI of .92, TLI of .90, RMSEA of .08, and SRMR of .06. Although these values suggest reasonably close fit, they did not meet all of the cutoff criteria established by Hu and Bentler (1999). The corresponding one-factor model performed considerably worse, with a CFI of .66, TLI of .61, RMSEA of .16, and SRMR of .11, indicating poor fit across all indices. The likelihood ratio test indicated that the five-factor model fit the data significantly better than the one-factor model, $\Delta\chi^2(10) = 347.11$, $p < .001$.

For the explanation-based scale, the five-factor model demonstrated strong fit, meeting all recommended cutoffs for model adequacy. Fit indices were CFI = .98, TLI = .98, RMSEA = .05, and SRMR = .05, all within the thresholds proposed by Hu and Bentler. In contrast, the one-factor model again exhibited poor fit, with a CFI of .63, TLI of .57, RMSEA of .18, and SRMR of .11. The likelihood ratio test indicated that the five-factor model fit the data significantly better than the one-factor model, $\Delta\chi^2(10) = 564.96$, $p < .001$.

Organizational Commitment Factor Structure. Table 10 presents model fit indices for CFAs conducted on organizational commitment scales generated by GPT using zero-shot, few-shot, and explanation-based prompting. Each scale was evaluated using one-factor, three-factor, and hierarchical factor structures. Based on theoretical foundations, I hypothesized that a three-factor structure, representing affective, continuance, and normative commitment, would demonstrate better fit compared to a one-factor solution, for the few-shot and explanation-based scales.

For the few-shot scale, the three-factor model showed mixed evidence of fit. Some indices approached conventional thresholds for acceptable fit, including a CFI of .90 and an SRMR of .08. However, others indicated poor fit, such as a TLI of .86 and an RMSEA of .18. The one-factor model demonstrated substantially worse fit, with a CFI of .80, TLI of .73, RMSEA of .24, and SRMR of .10. These results support the relative advantage of the three-factor solution. The likelihood ratio test indicated that the three-factor model fit the data significantly better than the one-factor model, $\Delta\chi^2(3) = 149.54$, $p < .001$.

For the explanation-based scale, the three-factor model exhibited strong overall fit. It met most of the cutoff criteria recommended by Hu and Bentler (1999), with a CFI of .97, TLI of .95, RMSEA of .09, and SRMR of .05. Although the RMSEA slightly exceeded the conventional

threshold of .06, the model's overall fit was strong. In contrast, the one-factor model performed poorly across all indices, with a CFI of .67, TLI of .56, RMSEA of .28, and SRMR of .18. The likelihood ratio test indicated that the three-factor model fit the data significantly better than the one-factor model, $\Delta\chi^2(3) = 379.76$, $p < .001$.

Organizational Justice Factor Structure. Table 11 presents model fit indices from CFAs conducted on organizational justice scales generated through zero-shot, few-shot, and explanation-based GPT prompting. Each scale was tested using one-factor, four-factor, and hierarchical models. I hypothesized that a three-factor solution would provide better fit than a one-factor model for both the few-shot and explanation-based scales, in line with prior conceptualizations of organizational justice dimensions.

For the few-shot scale, the four-factor model demonstrated strong fit, with values approaching or meeting conventional standards for model adequacy: CFI = .97, TLI = .96, RMSEA = .07, and SRMR = .03. These indices indicate that the model adequately captures the underlying structure of the data. In contrast, the one-factor model demonstrated poorer fit, with a CFI of .91, TLI of .90, RMSEA of .12, and SRMR of .05. The likelihood ratio test indicated that the four-factor model fit the data significantly better than the one-factor model, $\Delta\chi^2(6) = 100.84$, $p < .001$.

For the explanation-based scale, the four-factor model demonstrated excellent fit, meeting all of Hu and Bentler's (1999) recommended cutoffs. Fit indices were CFI = .99, TLI = .95, RMSEA = .06, and SRMR = .02, indicating a well-fitting model. The one-factor solution for this scale showed comparatively poor fit, despite some indices approaching acceptable levels: CFI = .90, TLI = .88, RMSEA = .06, and SRMR = .03. The likelihood ratio test indicated that

the four-factor model fit the data significantly better than the one-factor model, $\Delta\chi^2(6) = 188.99$, $p < .001$.

Nomological Network Validity Evidence

To evaluate the nomological network validity of the GPT-authored scales, correlations were examined between GPT-authored scales for job satisfaction, organizational commitment, and organizational justice and scales assessing theoretically related constructs using the full cleaned dataset (N = 346). See Table 12 for a full correlation matrix of variables within the nomological network.

Job Satisfaction Results. To demonstrate construct validity, Hypothesis 3a, which proposed that GPT-authored job satisfaction would be positively related to positive affect, extraversion, and conscientiousness, was supported. GPT-authored job satisfaction correlated positively with positive affect (zero-shot scale: $r = 0.52$, few-shot scale: $r = 0.50$, explanation-based scale: $r = 0.52$; $p < .05$), extraversion (zero-shot scale: $r = 0.18$, few-shot scale: $r = 0.19$, explanation-based scale: $r = 0.19$; $p < .05$), and conscientiousness (zero-shot scale: $r = 0.23$, few-shot scale: $r = 0.20$, explanation-based scale: $r = 0.22$; $p < .05$). Hypothesis 3b, which predicted negative relationships between GPT-authored job satisfaction and neuroticism, was supported (zero-shot scale: $r = -0.22$, few-shot scale: $r = -0.24$, explanation-based scale: $r = -0.26$; $p < .05$), as well as its negative relationship with negative affect (zero-shot scale: r = -0.30, few-shot scale: $r = -0.30$, explanation-based scale: $r = -0.31$; $p < .05$). These findings provide evidence for the construct validity of GPT-authored job satisfaction across prompting techniques. Consistent with Hypothesis 3c, GPT-authored job satisfaction positively correlated with organizational citizenship behaviors (OCB; zero-shot scale: $r = 0.25$, few-shot scale: $r = 0.25$, explanation-based scale: $r = 0.23$; $p < .05$). Supporting Hypothesis 3d, GPT-authored job satisfaction was

negatively related to counterproductive work behaviors (CWB; zero-shot scale: r = -0.19, few-shot scale: r = -0.18, explanation-based scale: r = -0.19; $p < .05$). These significant correlations provide evidence for the criterion-related validity of GPT-authored job satisfaction across prompts.

Organizational Commitment Results. Hypothesis 4a proposed that GPT-authored organizational commitment would be positively related to locus of control, extraversion, and conscientiousness. These relationships were supported, with GPT-authored organizational commitment correlating positively with locus of control (zero-shot scale: r = 0.15, few-shot scale: r = 0.17, explanation-based scale: r = 0.15; $p < .05$), extraversion (zero-shot scale: r = 0.16, few-shot scale: r = 0.18, explanation-based scale: r = 0.16; $p < .05$), and conscientiousness (zero-shot scale: r = 0.11, few-shot scale: r = 0.14, explanation-based scale: r = 0.11; $p < .05$). Hypothesis 4b, which predicted negative relationships between GPT-authored organizational commitment and both burnout and stress, was also supported. GPT-authored organizational commitment correlated negatively with burnout (zero-shot scale: r = -0.40, few-shot scale: r = -0.42, explanation-based scale: r = -0.39; $p < .05$) as well as job stress (zero-shot scale: r = -0.21, few-shot scale: r = -0.20, explanation-based scale: r = -0.19; $p < .05$).

Organizational Justice Results. Hypothesis 5a proposed that GPT-authored organizational justice would be positively related to positive affect and agreeableness. This was supported, with positive affect showing strong positive correlations (zero-shot scale: r = 0.48, few-shot scale: r = 0.46, explanation-based scale: r = 0.47; $p < .05$), while agreeableness exhibited weaker but significant positive correlations (zero-shot scale: r = 0.11, few-shot scale: r = 0.13, explanation-based scale: r = 0.11; $p < .05$). Hypothesis 5b predicted negative relationships between GPT-authored organizational justice and neuroticism, stress, and burnout.

As expected, GPT-authored organizational justice correlated negatively with negative affect (zero-shot scale: r = -0.26, few-shot scale: r = -0.26, explanation-based scale: r = -0.27; $p < .05$), neuroticism (zero-shot scale: r = -0.24, few-shot scale: r = -0.23, explanation-based scale: r = -0.23; $p < .05$), stress (zero-shot scale: r = -0.33, few-shot scale: r = -0.33, explanation-based scale: r = -0.34; $p < .05$), and burnout (zero-shot scale: r = -0.47, few-shot scale: r = -0.46, explanation-based scale: r = -0.48; $p < .05$). Supporting Hypothesis 5c, GPT-authored organizational justice positively correlated with organizational citizenship behaviors (OCB; zero-shot scale: r = 0.21, few-shot scale: r = 0.23, explanation-based scale: r = 0.23; $p < .05$). Similarly, Hypothesis 5d was supported, with GPT-authored organizational justice showing significant negative correlations with counterproductive work behaviors (CWB; zero-shot scale: r = -0.22, few-shot scale: r = -0.17, explanation-based scale: r = -0.21; $p < .05$).

Redundancy Results. Finally, research questions explored the relationship between GPT-authored and human-authored scales to evaluate redundancy. For job satisfaction, GPT-authored and human-authored measures showed high positive correlations (zero-shot scale: r = 0.80, few-shot scale: r = 0.79, explanation-based scale: r = 0.80; $p < .05$). Substantial correlations were also observed between GPT-authored and human-authored measures for organizational commitment (zero-shot scale: r = 0.78, few-shot scale: r = 0.78, explanation-based scale: r = 0.78; $p < .05$) and organizational justice (zero-shot scale: r = 0.90, few-shot scale: r = 0.88, explanation-based scale: r = 0.88; $p < .05$).

DISCUSSION

The field of I-O Psychology has increasingly incorporated LLMs into scale development methodologies (e.g., Hernandez & Nie, 2022). Given that LLMs respond to natural language, researchers have many options for prompting them to create scale items. The current study explored three distinct prompting techniques for item generation in the form of zero-shot, few-shot, and explanation-based approaches. Previous research has not systematically compared different prompting techniques and has primarily focused on the generation of Big Five personality items. This study aimed to evaluate the psychometric properties of GPT-authored scales measuring job satisfaction, organizational commitment, and organizational justice, generated through three different prompting techniques. To explore this, I evaluated the content validity, reliability, factor structure, and nomological network validity of each scale.

First, when evaluating content validity, the explanation-based scales received the highest ratings of definitional correspondence, followed closely by the zero-shot scales, while the few-shot scales performed the worst (see Table 5). Although few-shot prompting provides example items intended to guide the model, it may inadvertently reduce content validity due to semantic contamination, wherein the model interprets the specific content of examples as contextually meaningful in a way that constrains generalization (Reynolds & McDonell, 2021). Rather than treating examples as abstract representations of a construct category, the model may infer that their specific themes define the construct's scope. This can lead to generated items that narrowly reflect the example content rather than the broader conceptual definition.

Findings across factor analyses suggest that GPT-authored items, regardless of prompting technique, generally aligned with their intended theoretical constructs. This is particularly notable given that no human subject matter experts were involved in item writing, and all items

were drawn directly from GPT's initial output without iteration or selective refinement. Explanation-based prompts consistently produced the best-fitting models in CFAs across all constructs. In the context of scale development, providing explicit definitions of constructs and dimensions may help establish semantic boundaries, guiding the model to generate content that is more closely aligned with the intended latent structures. Without such definitions, GPT may rely on varied representations of a construct drawn from its pre-training data. By instructing the model to use definitions for item generation, explanation-based prompts "constrain" model behavior by reducing ambiguity and increasing the likelihood that items reflect the intended theoretical framework. Zero-shot prompts also performed surprisingly well. For these prompts, GPT was only given construct and dimension labels without definitions or examples. Despite this minimal input, the items generally loaded onto their intended factors, as supported by both EFA and CFA results. Zero-shot items for job satisfaction and organizational commitment aligned well with expected theoretical factor structures, indicating that the model represented these constructs in a conceptually consistent way. Organizational justice was a partial exception, as exploratory factor analysis revealed a three-factor solution rather than the expected four-factor structure. This overlap is not entirely unexpected, given prior research showing moderate correlations between procedural and distributive justice (r = .55) and between procedural and informational justice (r = .56; Colquitt & Shaw, 2005). However, CFAs in the second sub-sample showed that the four-factor model fit the data better than the three-factor model when comparing fit statistics (see Table 11), contrasting with the EFA results and supporting the theoretical expectation of a four-factor structure for organizational justice. Few-shot prompting generally resulted in the poorest model fit across constructs. Although all examples were drawn from validated scales, providing examples alone may have constrained the model's output by

encouraging GPT to overfit to the specific content of the examples rather than generalizing to the broader construct.

This aligns with prior research (Reynolds & McDonell, 2021) indicating that models may interpret few-shot examples as contextually meaningful in a way that constrains generalization, especially in the case that a lower number of examples are used. However, the organizational justice scale was a notable exception, showing relatively strong model fit under the few-shot prompt. This may be due to the more semantically distinct and easily separable language used in justice dimensions, such as “procedures,” “explanations,” and “respect,” which may have helped GPT maintain clearer conceptual boundaries.

The evaluation of nomological network validity revealed strong consistency between GPT-authored scales and established theoretical expectations. For job satisfaction, GPT-authored scales demonstrated robust positive correlations with positive affect (r = .50 to .52), closely aligning with prior meta-analytic findings (Thoresen et al., 2003; ρ = .34). Similarly, negative correlations with negative affect (r = -.30 to -.31) and neuroticism (r = -.22 to -.26) mirrored established trait-affect associations in the literature. Correlations with extraversion (r = .18 to .19) and conscientiousness (r = .20 to .23) were also directionally consistent with population estimates (Judge et al., 2002), reinforcing the construct validity of the GPT-authored scales. While the negative association between job satisfaction and counterproductive work behaviors (CWB; r = -.18 to -.19) was smaller than prior estimates (Dalal, 2005; ρ = -.37), the result remains theoretically meaningful and in the expected direction. Organizational commitment showed a similarly coherent nomological profile. It was positively related to locus of control, extraversion, and conscientiousness, and negatively associated with burnout and job stress, which is consistent with prior meta-analytic patterns (Choi et al., 2015). The nomological

network for organizational justice also aligned well with expectations. Positive associations with positive affect, agreeableness, and OCB, alongside negative relationships with neuroticism, stress, burnout, and CWB, reflect well-established patterns in the literature (Colquitt et al., 2012). Importantly, these relationships held consistently across all three prompting techniques (zero-shot, few-shot, and explanation-based).

Taken together, these findings provide strong evidence that GPT-authored items function in theoretically coherent ways. While the observed correlations are not always identical to meta-analytic estimates, the overall pattern is consistent with the nomological networks of validated measures. This supports the conclusion that large language models, when prompted effectively, can generate items that are not only psychometrically sound but also embedded in well-established theoretical frameworks.

Theoretical Contributions

This study makes several significant contributions to the AIG literature. First, this is the first study to place prompt engineering at the center of AIG research. Previous studies have not studied prompt engineering (von Davier, 2018; Hommel et al., 2021) or examined only a single approach (Lee et al., 2023). In contrast, this study identifies three prompting techniques from prompt engineering research: zero-shot, few-shot, and explanation-based prompting. These techniques are examined to compare their effectiveness in AIG. One novel approach in AIG is the use of explanations to supplement task instructions. In this study, I operationalized explanation-based prompting by providing the LLM with construct and dimension definitions to clarify and reinforce the intended meaning of the construct. The goal of the method was to enhance the model's ability to generate items that are both semantically precise and theoretically grounded. By explicitly defining the construct, it reduces ambiguity by narrowing the model's

generative space, ensuring that token selection aligns with the intended psychological framework and remains semantically consistent with the construct.

Second, this study is the first to generate items specifically for non-personality I-O psychology constructs, addressing a gap in AIG research. Prior studies have predominantly focused on personality, leaving an open question regarding the effectiveness of AI-generated items in measuring workplace attitudes. By demonstrating that GPT-authored items for job satisfaction, organizational commitment, and organizational justice exhibit strong psychometric properties, including reliability, validity, and factor structure, this study expands the scope of AIG applications beyond personality measurement.

Practical Contributions

This study has multiple practical contributions that can be framed through Thomas and Tymon's (1982) five criteria for relevant research: descriptive relevance, goal relevance, operational validity, non-obviousness, and timeliness.

Descriptive Relevance: Addressing Practitioner Interest

First, this study demonstrates descriptive relevance by addressing the growing interest among practitioners in effectively leveraging LLMs. The increasing prominence of AI-focused workshops, such as the Society for Industrial and Organizational Psychology (SIOP) 2024 Machine Learning Competition, along with related panels and symposia at professional conferences, underscores this trend. This study represents a step in integrating prompt engineering approaches from the artificial intelligence literature into I-O psychology.

Goal Relevance: Addressing Relevant Outcomes

Second, this study demonstrates goal relevance by evaluating the criterion-related validity of GPT-authored scales in relation to key workplace behaviors, specifically OCB and

CWB. By validating that GPT-authored scales can meaningfully predict OCB and CWB, this study provides additional evidence that regardless of whether a scale is written by a human or an LLM, it can be situated meaningfully within the nomological network of established work psychology measures.

Operational Validity: Accessible to Practitioners

Operational validity was established by using ChatGPT, a widely accessible and commonly used LLM, alongside prompting techniques that required neither extensive pre-existing item banks nor technical expertise. Effective item generation was achieved using simple, user-friendly prompts, such as providing only the construct label, a small set of example items, or a clear construct definition. The decision to study this explanation-based prompting technique stemmed from its alignment with how many users naturally interact with LLMs. Rather than relying on complex prompting strategies or highly specific use cases, this approach is a flexible and broadly applicable strategy, making it useful across a wide range of constructs.

Non-Obviousness: Contributing New Insights

There was a legitimate concern that zero-shot prompting might produce poorly performing items due to the lack of explicit guidance provided to the model. Moreover, since prior research had primarily focused on personality constructs, it was uncertain whether the model would perform equally well in generating items for non-personality constructs such as job satisfaction, organizational commitment, and organizational justice. Additionally, it was unclear whether providing a construct definition would meaningfully enhance item quality.

Zero-shot prompting performed well, successfully generating psychometrically valid scales despite the absence of explicit guidance or examples. While explanation-based prompting had the potential to introduce unnecessary complexity, it ultimately enhanced validity by

enabling the model to produce more precise and conceptually aligned items. This finding underscores the potential of both zero-shot and explanation-based prompting as viable techniques alongside few-shot prompting, expanding the range of effective approaches for LLM-assisted scale development.

Timeliness: Responsive to Emerging Issues

Finally, the timeliness of this study is reflected in the use of ChatGPT, the most advanced LLM available at the time of research. As LLMs continue to evolve, results based on a particular model may not fully generalize to future iterations, presenting an inherent challenge to this research. This study captures a critical moment, but future advances in LLMs may alter the nature of both outputs and the prompting processes themselves, stressing the need for ongoing validation. This inherent limitation of LLM research is discussed further in the section below, highlighting the need for ongoing evaluation as technology advances.

Limitations

The first limitation concerns generalizability in two areas: (1) across different models or future iterations of LLMs, and (2) across psychological constructs. Since the completion of this study, OpenAI has released an updated version of ChatGPT (GPT-4o), and ongoing advancements in LLM capabilities are expected. As these models continue to evolve, the prompting techniques used in this study may need to be adjusted to remain effective. For example, future LLMs may offer enhanced reasoning abilities, improved contextual understanding, or greater responsiveness to user intent. Such changes could influence both the nature and quality of the generated items, potentially altering how different prompting techniques shape model output. In terms of construct generalizability, this study builds on prior work that has primarily focused on personality by applying LLMs to job attitudes. This broader

application is a strength of the current research. However, it remains an open question how LLMs will perform when asked to generate items for constructs that are not yet well established in the academic literature.

Another limitation is the syntactic structure of items produced. Many items were double-barreled, such as “The pay raises and bonuses in my organization are fair and based on job performance” produced by the zero-shot prompt for job satisfaction or “In my organization, individuals are treated with respect and dignity by their superiors” generated by the explanation-based prompt for interactional justice. Double-barreled items were produced across prompts, although items generated by few-shot prompts tended to follow the syntactic structure of the example items more closely.

An important final limitation is the potential for plagiarized items. I used the software Turnitin to evaluate items for plagiarism. Turnitin is an online tool widely used for detecting plagiarism. It compares submitted documents against an extensive database of academic papers, online content, and previously submitted work to identify overlapping text. Across all constructs, dimensions, and items generated, three affective commitment items were flagged by Turnitin for plagiarism: (1) “I feel a strong sense of belonging to my organization,” (2) “I am proud to tell others that I am a part of this organization,” and (3) “I feel emotionally attached to this organization.” Researchers who use ChatGPT should use Turnitin or other tools to identify items that are potentially plagiarized.

Future Directions

Future research should explore a broader range of prompting techniques. In this study, the explanation-based approach was intentionally kept simple; however, there is considerable potential to expand on this method by incorporating additional “constraints” to model behavior

to guide item generation more precisely. Beyond providing construct definitions, future prompts could include explanations about the purpose of the scale, the target population, preferred syntactic structure, and other relevant specifications to produce more targeted items. Prompting techniques that allow LLMs to break up the problem in multiple steps, such as chain-of-thought prompting (Wei et al., 2022) or that pair examples with explanations (Lampinen et al., 2022) should also be explored.

It remains an open question how effectively LLMs can generate items for novel or emerging psychological constructs. These constructs may be underrepresented in the model's training corpus or discussed primarily in informal, non-academic contexts, which could limit the relevance, precision, or theoretical alignment of the generated items. In some cases, the model may rely on adjacent or vaguely related concepts, introducing construct-irrelevant variance. Future work should examine how model output quality varies as a function of construct maturity and how additional guidance or prompting techniques might mitigate these issues.

Another important consideration is the potential for plagiarism in LLM-authored items. Although this study made efforts to address the possibility of plagiarism in generated items using the tool Turnitin, future research should more systematically investigate various approaches and tools for effectively identifying plagiarized items. Establishing best practices for evaluating item originality will be essential to promoting responsible and transparent use of LLMs in psychological research.

Conclusion

The findings suggest that ChatGPT demonstrates considerable potential for generating scale items, supported by evidence of content validity, construct validity, and criterion-related validity. Among the three prompting techniques tested, explanation-based prompting was the

most effective, followed closely by zero-shot prompting, while few-shot prompting yielded weaker results. These findings highlight the promise of GPT-based tools for scale development and emphasize the importance of prompt engineering in improving performance.

Table 1

Content for Prompts (Construct Labels, Example Items, Definition)

Construct Label	Example Item	Definition (Explanation)
Job Satisfaction		
Supervisor	My supervisor praises good work (Stanton et al., 2002)	Degree to which the employee is satisfied with how their supervisor supervises and/or the qualities the supervisor possesses (Job Descriptive Index; Stanton et al., 2002).
Coworker	My coworkers are helpful (Stanton et al., 2002)	Degree to which employees view their coworkers as having positive/negative qualities or behavioral tendencies (Job Descriptive Index; Stanton et al., 2002).
Pay	My job pay is fair (Stanton et al., 2002)	Degree to which an employee perceives their job's monetary compensation as fair and adequate (Job Descriptive Index; Stanton et al., 2002).
Promotional	I have a good chance for promotion (Stanton et al., 2002)	Degree to which the employee believes the job role offers opportunities for fair job advancement (Job Descriptive Index; Stanton et al., 2002).
Work	My job gives me a sense of accomplishment (Stanton et al., 2002)	Degree to which an employee finds their job and the tasks they perform satisfying, challenging, and/or interesting (Job Descriptive Index; Stanton et al., 2002).
Organizational Commitment		
Affective	This organization has a great deal of personal meaning to me (Meyer et al. 1993)	Degree to which an employee feels a sense of belongingness to their organization, feels emotionally attached to their organization, and/or feels personally connected to their organization. (Choi et al., 2015; Meyer et al., 1993).
Normative	This organization deserves my loyalty (Meyer et al. 1993)	Commitment based on felt obligation to remain with the organization (Choi et al., 2015; Meyer et al., 1993).
Continuance	I feel that I have too few options to consider leaving this organization (Meyer et al. 1993)	Commitment based on the costs and benefits associated with leaving the organization (Choi et al., 2015; Meyer et al., 1993).
Organizational Justice		
Distributive	Does your (outcome) reflect the effort you have put into your work? (Colquitt et al., 2001)	Refers to the perceived fairness of outcome distributions (Ambrose et al., 2007).
Procedural	Have those procedures been applied consistently? (Colquitt et al., 2001)	Refers to the perceived justice of the process by which decisions are made (Ambrose et al., 2007).
Informational	Has (he/she) explained the procedures thoroughly? (Colquitt et al., 2001)	Refers to explanations one receives during the enactment of procedures. Specifically, the timeliness, completeness, and accuracy of the information received in an organization (Ambrose et al., 2007; Greenberg, 1993).
Interactional	Has (he/she) treated you with respect? (Colquitt et al., 2001)	Refers to interpersonal sensitivity and the fairness with how people are treated within an organization. (Ambrose et al., 2007; Greenberg, 1993).

Note: Table adapted from Speer et al. (2022)

Table 2

Content Validity of GPT-Authored Job Satisfaction Scale

Factor	Prompt	Item	htc
Supervisor	Zero-Shot	My supervisor gives me the support and guidance I need to do my job effectively.	0.90
Supervisor	Zero-Shot	I receive constructive feedback from my supervisor that helps me improve my performance.	0.90
Supervisor	Zero-Shot	My supervisor recognizes and appreciates my contributions to the team.	0.87
Supervisor	Few-Shot	My supervisor provides constructive feedback.	0.84
Supervisor	Few-Shot	My supervisor effectively communicates expectations.	0.86
Supervisor	Few-Shot	My supervisor is accessible when I need support or guidance.	0.85
Supervisor	Explanation	I feel that my supervisor values my input and contributions at work.	0.88
Supervisor	Explanation	My supervisor provides me with constructive feedback that helps me grow professionally.	0.89
Supervisor	Explanation	I am satisfied with the level of communication and openness I have with my supervisor.	0.87
Pay	Zero-Shot	I am satisfied with my current level of pay compared to the demands of my job.	0.82
Pay	Zero-Shot	My pay reflects the amount of effort I put into my work.	0.77
Pay	Zero-Shot	The pay raises and bonuses in my organization are fair and based on job performance.	0.81
Pay	Few-Shot	My salary reflects the responsibilities of my job.	0.78
Pay	Few-Shot	I believe my pay is competitive with similar roles in other organizations.	0.79
Pay	Few-Shot	The pay raise process in my organization is transparent and fair.	0.73
Pay	Explanation	I believe my pay reflects the effort and quality of work I bring to my job.	0.80
Pay	Explanation	Compared to similar roles in my industry, I feel that I am compensated fairly.	0.81
Pay	Explanation	My salary/benefits package is competitive and meets my financial needs.	0.79
Coworker	Zero-Shot	My coworkers are supportive and help create a positive work environment.	0.90
Coworker	Zero-Shot	I feel like I am part of a team and can rely on my colleagues when needed.	0.84
Coworker	Zero-Shot	There is a good sense of camaraderie among the employees in my department.	0.82
Coworker	Few-Shot	My coworkers are committed to doing quality work.	0.78
Coworker	Few-Shot	My coworkers support a positive work environment.	0.88
Coworker	Few-Shot	I trust my coworkers and work well with them.	0.84
Coworker	Explanation	I believe my coworkers are supportive and help create a positive work environment.	0.88
Coworker	Explanation	My coworkers and I collaborate effectively on projects and tasks.	0.82
Coworker	Explanation	I feel respected and valued by my coworkers.	0.86
Promotion	Zero-Shot	Opportunities for advancement are provided equally to all employees in my organization.	0.82
Promotion	Zero-Shot	I am satisfied with my prospects for career advancement within this organization.	0.85
Promotion	Zero-Shot	The criteria for promotions are clear and transparent to everyone.	0.77
Promotion	Few-Shot	The criteria for promotion are clear and applied consistently.	0.78
Promotion	Few-Shot	Promotions are frequent enough to motivate employees.	0.77
Promotion	Few-Shot	The organization supports my career development and advancement opportunities.	0.82
Promotion	Explanation	I see clear and fair opportunities for career advancement within my organization.	0.87
Promotion	Explanation	The criteria for promotion are transparent and applied consistently to all employees.	0.80
Promotion	Explanation	I feel motivated by the career development opportunities available to me here.	0.78
Work	Zero-Shot	The work I do is meaningful and gives me a sense of accomplishment.	0.87
Work	Zero-Shot	I am given enough freedom to decide how to best perform my work.	0.70
Work	Zero-Shot	My job provides me with a variety of tasks that challenge and engage me.	0.86
Work	Few-Shot	My job provides opportunities for personal growth and learning.	0.77
Work	Few-Shot	My job tasks are varied and interesting.	0.83
Work	Few-Shot	I feel my job contributes positively to the organization's goals.	0.72
Work	Explanation	I find the tasks and responsibilities of my job to be engaging and meaningful.	0.88
Work	Explanation	My job offers a good balance of challenges and opportunities for learning.	0.83
Work	Explanation	I feel that my job allows me to fully utilize my skills and abilities.	0.82

Table 3

Content Validity of GPT-Authored Organizational Commitment Scale

Factor	Prompt	Item	htc
Affective	Zero-Shot	I feel a strong sense of belonging to my organization.	0.87
Affective	Zero-Shot	I am proud to tell others that I am part of this organization.	0.82
Affective	Zero-Shot	Working for this organization gives me a great sense of personal accomplishment.	0.80
Affective	Few-Shot	Seeing my organization's success feels like a personal achievement to me.	0.81
Affective	Few-Shot	I feel emotionally attached to this organization.	0.84
Affective	Few-Shot	The values of this organization deeply resonate with my personal values.	0.80
Affective	Explanation	I feel a strong sense of belonging to my organization.	0.89
Affective	Explanation	I am proud to tell others I am part of this organization.	0.81
Affective	Explanation	My organization's successes feel like my own.	0.83
Normative	Zero-Shot	I feel a moral obligation to remain with my current employer.	0.80
Normative	Zero-Shot	Even if it were to my advantage, leaving my organization would feel wrong.	0.77
Normative	Zero-Shot	I feel a strong sense of loyalty to my colleagues, which encourages me to stay with this organization.	0.81
Normative	Few-Shot	I believe in the ethos of sticking with one's organization through good times and bad.	0.75
Normative	Few-Shot	Leaving this organization would feel like betraying the trust people have in me.	0.77
Normative	Few-Shot	The values of this organization deeply resonate with my personal values.	0.62
Normative	Explanation	I feel that I ought to remain with my current organization even if it were not in my best interest to stay.	0.76
Normative	Explanation	Leaving my organization now would feel like betraying my colleagues.	0.74
Normative	Explanation	I feel an obligation to continue working for my organization.	0.82
Continuance	Zero-Shot	I feel that I have too much to lose by leaving my organization right now.	0.73
Continuance	Zero-Shot	It would be hard for me to leave my organization because I would lose many valuable benefits.	0.66
Continuance	Zero-Shot	I feel that my options are limited right now, which makes staying at my organization necessary.	0.57
Continuance	Few-Shot	The investment I have made in this organization makes it difficult for me to think about leaving.	0.75
Continuance	Few-Shot	I feel that the time and effort I have put into this organization bond me to it.	0.77
Continuance	Few-Shot	Leaving this organization would mean losing a lot of what I've built here, such as relationships and knowledge.	0.75
Continuance	Explanation	Staying with my organization is a practical necessity for me.	0.66
Continuance	Explanation	I am staying with my organization because of what I would lose if I left.	0.70
Continuance	Explanation	The thought of leaving my organization seems unwise to me because of the investments I have made in it.	0.74

Table 4

Content Validity of GPT-Authored Organizational Justice Scale

Factor	Prompt	Item	htc
Distributive	Zero-Shot	I feel that the rewards I receive (e.g., pay, recognition) are proportional to the effort I put into my work.	0.77
Distributive	Zero-Shot	The workload is distributed fairly among employees in my department.	0.80
Distributive	Zero-Shot	My organization ensures that employees receive fair compensation for the contributions they make.	0.82
Distributive	Few-Shot	Do you feel the distribution of tasks and responsibilities in your workplace is fair?	0.72
Distributive	Few-Shot	Are the criteria for performance evaluations clearly communicated and justly applied?	0.67
Distributive	Few-Shot	Does the distribution of promotions and raises within your organization seem equitable?	0.72
Distributive	Explanation	I believe the outcomes of decisions in my organization are distributed fairly among employees.	0.81
Distributive	Explanation	The reward system in my organization reflects the effort and performance of the employees accurately.	0.78
Distributive	Explanation	In my organization, promotions are given based on merit and are fair to all employees.	0.82
Procedural	Zero-Shot	Decisions regarding promotions are made through transparent and impartial procedures.	0.80
Procedural	Zero-Shot	I believe that the policies and procedures that affect me are applied consistently.	0.76
Procedural	Zero-Shot	My organization allows employees to voice their concerns before making decisions that affect them."	0.76
Procedural	Few-Shot	Do decision-making processes at work seem fair and impartial to you?	0.74
Procedural	Few-Shot	Are you informed about how decisions that affect you are made?	0.64
Procedural	Few-Shot	Is there a mechanism in place for appealing or challenging decisions made at work?	0.68
Procedural	Explanation	The processes used for making decisions in my organization are applied consistently across all cases.	0.81
Procedural	Explanation	I feel that the procedures for decision making in my organization allow for all voices to be heard.	0.84
Procedural	Explanation	The methods used to resolve disputes in my organization are impartial and just.	0.81
Informational	Zero-Shot	I am provided with adequate explanations about decisions that affect my work.	0.80
Informational	Zero-Shot	The organization communicates changes in policies or procedures in a timely and clear manner.	0.86
Informational	Zero-Shot	I receive constructive feedback that helps me improve my performance.	0.71
Informational	Few-Shot	Are you kept informed about important issues and changes within the organization in a timely manner?	0.77
Informational	Few-Shot	Does the organization provide adequate explanations for decisions and actions that affect you?	0.73
Informational	Few-Shot	Is feedback provided in a constructive and helpful manner, allowing for personal and professional growth?	0.69
Informational	Explanation	I am satisfied with the timeliness of the information provided by my organization during decision-making processes.	0.83
Informational	Explanation	The information I receive from my organization about decisions is complete and comprehensible.	0.82
Informational	Explanation	My organization provides accurate and truthful explanations regarding decisions and procedures.	0.84
Interactional	Zero-Shot	My supervisor communicates with me in a respectful manner.	0.86
Interactional	Zero-Shot	Feedback, whether positive or negative, is delivered to me in a considerate and tactful manner.	0.81
Interactional	Zero-Shot	I feel that my opinions are valued and respected by management.	0.79
Interactional	Few-Shot	Are you treated with kindness and understanding by your supervisors?	0.78
Interactional	Few-Shot	Do your colleagues treat you in a manner that makes you feel valued and respected?	0.77
Interactional	Few-Shot	Does the management acknowledge your personal circumstances and treat you accordingly?	0.75
Interactional	Explanation	In my organization, individuals are treated with respect and dignity by their superiors.	0.87
Interactional	Explanation	Feedback in my organization is delivered in a considerate and constructive manner.	0.78
Interactional	Explanation	My superiors in the organization communicate in a way that makes me feel valued and respected.	0.84

Table 5

Average htc across constructs

Construct	Zero-shot	Few-shot	Explanation
Job Satisfaction	0.83	0.80	0.84
Org. Commit.	0.76	0.76	0.77
Org. Justice	0.80	0.72	0.82

Table 6

EFA Factor Loadings - GPT-Authored Job Satisfaction

Factor	Item	Loading
Supervisor	Item 1	0.74
Supervisor	Item 2	1.12
Supervisor	Item 3	0.58
Pay	Item 1	0.75
Pay	Item 2	1.09
Pay	Item 3	0.65
Coworker	Item 1	1.02
Coworker	Item 2	0.83
Coworker	Item 3	0.83
Promotion	Item 1	0.78
Promotion	Item 2	0.86
Promotion	Item 3	0.84
Work	Item 1	0.46
Work	Item 2	0.28
Work	Item 3	1.18

Table 7

EFA Factor Loadings - GPT-Authored Organizational Commitment

Factor	Item	Loading
Affective	Item 1	0.88
Affective	Item 2	0.83
Affective	Item 3	1.03
Normative	Item 1	0.89
Normative	Item 2	1.03
Normative	Item 3	0.73
Continuance	Item 1	0.85
Continuance	Item 2	0.52
Continuance	Item 3	0.52

Table 8

EFA Factor Loadings - GPT-Authored Organizational Justice

Factor #	Factor	Item	Loading
1	Distributive	Item 1	1.09
1	Distributive	Item 3	0.73
1	Procedural	Item 1	0.62
2	Interactional	Item 1	1.01
2	Interactional	Item 2	0.78
2	Interactional	Item 3	0.70
3	Informational	Item 1	0.90
3	Informational	Item 2	0.91
3	Informational	Item 3	0.67
3	Procedural	Item 2	0.42
3	Procedural	Item 3	0.73

Table 9

GPT-Authored Job Satisfaction CFA Model Fit

Prompt	Models	$\chi 2$	df	p	CFI	TLI	RMSEA	SRMR
Zero-shot	Five-Factor	160.69	80	<.01	.95	.93	.07 [.06, .09]	.06
	One-Factor	642.99	90	<.01	.63	.57	.18 [.16, .19]	.11
	Hierarchical	173.91	85	<.01	.94	.93	.07 [.06, .09]	.07
Few-Shot	Five-Factor	181.70	80	<.01	.92	.90	.08 [.07, .10]	.06
	One-Factor	528.81	90	<.01	.66	.61	.16 [.15 .17]	.11
	Hierarchical	195.13	85	<.01	.92	.90	.08 [.07, .10]	.07
Explanation-based	Five-Factor	112.02	80	<.01	.98	.98	.05 [.02, .06]	.05
	One-Factor	642.95	90	<.01	.63	.57	.18 [.16, .19]	.11
	Hierarchical	126.86	85	<.01	.98	.97	.05 [.03, .07]	.05

Note. Bracketed RMSEA values correspond to a 90% confidence interval at p = 05.

Table 10

GPT-Authored Organizational Commitment CFA Model Fit

Prompt	Models	$\chi 2$	df	p	CFI	TLI	RMSEA	SRMR
Zero-shot	Three-Factor	58.74	24	<.01	.97	.96	.09 [.06, .11]	.05
	One-Factor	399.24	27	<.01	.71	.61	.27 [.24, .29]	.14
	Hierarchical	47.73	24	<.01	.99	.99	.07 [.04, .10]	.05
Few-Shot	Three-Factor	180.90	24	<.01	.90	.86	.18 [.16, .21]	.08
	One-Factor	330.44	27	<.01	.80	.73	.24 [.22, .26]	.10
	Hierarchical	161.95	24	<.01	.99	.99	.17 [.15, .20]	.08
Explanation-based	Three-Factor	64.76	24	<.01	.97	.95	.09 [.07, .12]	.05
	One-Factor	444.51	27	<.01	.67	.56	.28 [.26, .30]	.18
	Hierarchical	39.91	24	.02	.99	.99	.06 [.02, .09]	.05

Note. Bracketed RMSEA values correspond to a 90% confidence interval at p = .05

Table 11

GPT-Authored Organizational Justice CFA Model Fit

Prompt	Models	$\chi 2$	df	p	CFI	TLI	RMSEA	SRMR
Zero-shot	Four-Factor	133.46	48	<.01	.96	.94	.10 [.08, .12]	.05
	Three-Factor	153.63	41	<.01	.94	.91	.12 [.10, .14]	.05
	One-Factor	269.26	54	<.01	.89	.86	.14 [.13, .16]	.06
	Hierarchical	153.47	50	<.01	.95	.93	.10 [.08, .12]	.05
Few-Shot	Four-Factor	95.28	48	<.01	.97	.96	.07 [.05, .09]	.03
	One-Factor	196.13	54	<.01	.91	.90	.12 [.10, .13]	.05
	Hierarchical	107.10	50	<.01	.97	.95	.08 [.06, .10]	.04
Explanation-based	Four-Factor	80.90	48	<.01	.99	.95	.06 [.04, .08]	.02
	One-Factor	269.84	54	<.01	.90	.88	.14 [.13, .16]	.05
	Hierarchical	93.22	50	<.01	.98	.97	.06 [.05, .09]	.03

Note. Bracketed RMSEA values correspond to a 90% confidence interval at p = .05.

Table 12

Full Correlation Matrix of Nomological Network Variables

Variable	Mean	SD	1	2	3	4	5	6	7	8
1. Zero-Shot Job Satisfaction	2.33	0.53								
2. Few-Shot Job Satisfaction	2.34	0.51	0.95***							
3. Explanation-Based Job Satisfaction	2.31	0.55	0.95***	0.94***						
4. Zero-Shot Org. Commitment	4.31	1.35	0.61***	0.58***	0.63***					
5. Few-Shot Org. Commitment	4.17	1.52	0.63***	0.61***	0.66***	0.92***				
6. Explanation-Based Org. Commitment	4.20	1.36	0.60***	0.57***	0.62***	0.96***	0.93***			
7. Zero-Shot Org. Justice	3.10	0.98	0.80***	0.80***	0.80***	0.58***	0.62***	0.57***		
8. Few-Shot Org. Justice	3.12	0.95	0.78***	0.79***	0.79***	0.59***	0.63***	0.58***	0.96***	
9. Explanation-Based Org. Justice	3.04	1.00	0.78***	0.79***	0.78***	0.58***	0.62***	0.57***	0.96***	0.95***
10. Job Satisfaction	2.28	0.34	0.80***	0.79***	0.80***	0.56***	0.58***	0.54***	0.72***	0.71***
11. Org. Commitment	4.07	0.87	0.50***	0.48***	0.51***	0.78***	0.78***	0.78***	0.51***	0.50***
12. Org. Justice	3.17	0.98	0.77***	0.76***	0.77***	0.62***	0.66***	0.61***	0.90***	0.88***
13. Positive Affect	4.13	0.94	0.52***	0.50***	0.52***	0.44***	0.48***	0.45***	0.48***	0.46***
14. Negative Affect	2.95	0.82	-0.30***	-0.30***	-0.31***	-0.18***	-0.22***	-0.19***	-0.26***	-0.26***
15. Extraversion	3.71	0.88	0.18***	0.19***	0.19***	0.16**	0.18***	0.16**	0.18***	0.18**
16. Agreeableness	5.26	1.05	0.09	0.1	0.07	0.19***	0.15**	0.17**	0.11*	0.13*
17. Conscientiousness	5.01	1.14	0.23***	0.20***	0.22***	0.11*	0.14**	0.11*	0.20***	0.19***
18. Neuroticism	3.79	1.26	-0.22***	-0.24***	-0.26***	-0.15**	-0.18***	-0.16**	-0.24***	-0.23***
19. Openness to Experience	5.16	1.11	-0.04	-0.03	-0.04	0	-0.02	0	-0.04	-0.03
20. Burnout	4.02	1.94	-0.56***	-0.51***	-0.54***	-0.40***	-0.42***	-0.39***	-0.47***	-0.46***
21. Job Stress	3.85	2.35	-0.37***	-0.33***	-0.35***	-0.21***	-0.20***	-0.19***	-0.33***	-0.33***
22. OCB	1.90	0.18	0.25***	0.25***	0.23***	0.27***	0.28***	0.27***	0.21***	0.23***
23. CWB	1.30	0.28	-0.19***	-0.18***	-0.19***	-0.19***	-0.20***	-0.21***	-0.22***	-0.17**
24. Locus of Control	4.82	0.70	0.18***	0.15**	0.16**	0.15**	0.17**	0.15**	0.17**	0.16**

Note. N = 346, *** p< .001, ** p < .01, * p<.05. This table continues onto the next page.

Table 12 (Continued)

Full Correlation Matrix of Nomological Network Variables

Variable	Mean	SD	9	10	11	12	13	14	15	16
9. Explanation-Based Org. Justice	3.04	1.00								
10. Job Satisfaction	2.28	0.34	0.69***							
11. Org. Commitment	4.07	0.87	0.49***	0.50***						
12. Org. Justice	3.17	0.98	0.88***	0.72***	0.55***					
13. Positive Affect	4.13	0.94	0.47***	0.52***	0.43***	0.51***				
14. Negative Affect	2.95	0.82	-0.27***	-0.39***	-0.1	-0.24***	-0.30***			
15. Extraversion	3.71	0.88	0.19***	0.1	0.14**	0.16**	0.36***	-0.09		
16. Agreeableness	5.26	1.05	0.11*	0.20***	0.22***	0.14*	0.12*	-0.16**	0.13*	
17. Conscientiousness	5.01	1.14	0.23***	0.23***	0.11*	0.22***	0.29***	-0.38***	0.14*	0.15**
18. Neuroticism	3.79	1.26	-0.23***	-0.30***	-0.07	-0.25***	-0.38***	0.53***	-0.28***	-0.13*
19. Openness to Experience	5.16	1.11	-0.04	0.02	-0.01	-0.03	0.1	-0.05	0.1	0.39***
20. Burnout	4.02	1.94	-0.48***	-0.58***	-0.31***	-0.49***	-0.43***	0.43***	-0.13*	-0.1
21. Job Stress	3.85	2.35	-0.34***	-0.48***	-0.11*	-0.34***	-0.22***	0.43***	0.01	-0.07
22. OCB	1.90	0.18	0.23***	0.25***	0.25***	0.23***	0.20***	-0.23***	0.06	0.24***
23. CWB	1.30	0.28	-0.21***	-0.24***	-0.16**	-0.20***	-0.30***	0.23***	-0.04	-0.1
24. Locus of Control	4.82	0.70	0.19***	0.20***	0.17**	0.17**	0.19***	-0.08	0.04	0.20***

Note. N = 346, *** p< .001, ** p < .01, * p<.05. This table continues onto the next page.

Table 12 (Continued)

Full Correlation Matrix of Nomological Network Variables

Variable	Mean	SD	17	18	19	20	21	22	23	24
17. Conscientiousness	5.01	1.14								
18. Neuroticism	3.79	1.26	-0.43***							
19. Openness to Experience	5.16	1.11	-0.02	-0.07						
20. Burnout	4.02	1.94	-0.23***	0.29***	-0.02					
21. Job Stress	3.85	2.35	-0.19***	0.27***	-0.02	0.68***				
22. OCB	1.90	0.18	0.17**	-0.15**	0.08	-0.19***	-0.13*			
23. CWB	1.30	0.28	-0.25***	0.29***	-0.14*	0.23***	0.14**	-0.18***		
24. Locus of Control	4.82	0.70	0.05	0.14**	0.45***	-0.13*	-0.04	0.18***	-0.12*	

Note. N = 346, *** p< .001, ** p < .01, * p<.05

REFERENCES

Ambrose, M., Hess, R. L., & Ganesan, S. (2007). The relationship between justice and attitudes: An examination of justice effects on event and system-related attitudes. *Organizational Behavior and Human Decision Processes*, *103*(1), 21–36. https://doi.org/10.1016/j.obhdp.2007.01.001

Baghaee, T. (2017). *Automatic neural question generation using community-based question answering systems* (Publication No. 10745163) [Masters Thesis, University of Lethbridge]. ProQuest Dissertations & Theses.

Boswell, W. R., Olson-Buchanan, J. B., & LePine, M. A. (2004). Relations between stress and work outcomes: The role of felt challenge, job control, and psychological strain. *Journal of Vocational Behavior*, *64*(1), 165–181. https://doi.org/10.1016/S0001-8791(03)00049-6

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & Agarwal, S. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Cachon-Alonso, L., & Elovainio, M. (2022). Organizational justice and health: reviewing two decades of studies. *Journal of theoretical social psychology*, *2022*(1), 3218883. https://doi.org/10.1155/2022/3218883

Choi, D., Oh, I. S., & Colbert, A. E. (2015). Understanding organizational commitment: A meta-analytic examination of the roles of the five-factor model of personality and culture. *Journal of applied psychology*, *100*(5), 1542. https://doi.org/10.1037/apl0000014

Colquitt, J. A., Conlon, D. E., Wesson, M. J., Porter, C. O., & Ng, K. Y. (2001). Justice at the millennium: A meta-analytic review of 25 years of organizational justice research.

Journal of Applied Psychology, *86*(3), 425–445. https://doi.org/10.1037/0021-9010.86.3.425

Colquitt, J. A., Scott, B. A., Rodell, J. B., Long, D. M., Zapata, C. P., Conlon, D. E., & Wesson, M. J. (2013). Justice at the millennium, a decade later: A meta-analytic test of social exchange and affect-based perspectives. *Journal of Applied Psychology*, *98*(2), 199–236. https://doi.org/10.1037/a0031757

Colquitt, J. A., & Shaw, J. C. (2005). How should organizational justice be measured? In J. Greenberg & J. A. Colquitt (Eds.), *Handbook of organizational justice* (pp. 113–152). Taylor & Francis Group.

Dalal, R. S. (2005). A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. *Journal of Applied Psychology*, *90*(6), 1241–1255. https://doi.org/10.1037/0021-9010.90.6.1241

Dalal, R. S., Lam, H., Weiss, H. M., Welch, E. R., & Hulin, C. L. (2009). A within-person approach to work behavior and performance: Concurrent and lagged citizenship-counterproductivity associations, and dynamic relationships with affect and overall job performance. *Academy of Management Journal*, *52*(5), 1051–1066. https://doi.org/10.5465/AMJ.2009.44636148

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment*, *18*(2), 192–203. https://doi.org/10.1037/1040-3590.18.2.192

Efrat, A., & Levy, O. (2020). The turking test: Can language models understand instructions? *arXiv Preprint*. https://doi.org/10.48550/arXiv.2010.11982

Erdheim, J., Wang, M., & Zickar, M. J. (2006). Linking the Big Five personality constructs to organizational commitment. *Personality and individual differences*, *41*(5), 959-970. https://doi.org/10.1016/j.paid.2006.04.005

Gao, T., Fisch, A., & Chen, D. (2021). Making pre-trained language models better few-shot learners. *arXiv Preprint*. https://doi.org/10.48550/arXiv.2012.15723

Gierl, M. J., Lai, H., & Tanygin, V. (2021). *Advanced Methods in Automatic Item Generation*. Routledge. https://doi.org/10.4324/9781003025634

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe, 7*(1), 7–28.

Greenberg, J. (1993). Organizational justice: Yesterday, today, and tomorrow. *Journal of Management*, *16*(2), 399–432. https://doi.org/10.1177/014920639001600208

Hackney, C. W. (2012). *Personality, organizational commitment, and job search behavior: a field study*. (Doctoral dissertation, The University of Tennessee).

Hase, P., & Bansal, M. (2021). When can models learn from explanations? A formal framework for understanding the roles of explanation data. *arXiv Preprint*. https://doi.org/10.48550/arXiv.2102.02201

Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But…. *Communication Methods and Measures*, *14*(1), 1-24. https://doi.org/10.1080/19312458.2020.1718629

Hernandez, I., & Nie, W. (2022). The AI-IP: Minimizing the guesswork of personality scale item development through artificial intelligence. *Personnel Psychology*, *76*(4), 1011–1035. https://doi.org/10.1111/peps.12543

Hinkin, T. R., & Tracey, J. B. (1999). An analysis of variance approach to content validation. *Organizational Research Methods*, *2*(2), 175–186. https://doi.org/10.1177/109442819922004

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/10.1016/0149-2063(95)90050-0

Hommel, B. E., Wollang, F. J. M., Kotova, V., Zacher, H., & Schmukle, S. C. (2021). Transformer-based deep neural language modeling for construct-specific automatic item generation. *Psychometrika*, *87*(2), 749–772. https://doi.org/10.1007/s11336-021-09823-9

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. https://doi.org/10.1007/BF02289447

Hu, G. (2023). Challenges for enforcing editorial policies on AI-generated papers. *Accountability in Research*, *31*(7), 978-980. https://doi.org/10.1080/08989621.2023.2184262

Hu, K. (2023, February 1). *ChatGPT sets record for fastest-growing user base – Analyst note*. Reuters. Retrieved March 21, 2023, from https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Johnson, R. E., Chang, C.-H. (Daisy), & Yang, L.-Q. (2010). Commitment and motivation at work: The relevance of employee identity and regulatory focus. *The Academy of Management Review*, *35*(2), 226–245. http://www.jstor.org/stable/25682410

Judge, T. A., Heller, D., & Mount, M. K. (2002). Five-factor model of personality and job satisfaction: A meta-analysis. *Journal of Applied Psychology*, *87*(3), 530–541. https://doi.org/10.1037/0021-9010.87.3.530

Judge, T. A., Zhang, S. C., & Glerum, D. R. (2020). Job satisfaction. *Essentials of job attitudes and other workplace psychological constructs*, 207-241.

Kinicki, A. J., McKee-Ryan, F. M., Schriesheim, C. A., & Carson, K. P. (2002). Assessing the construct validity of the job descriptive index: A review and meta-analysis. *Journal of Applied Psychology*, *87*(1), 14–32. https://doi.org/10.1037/0021-9010.87.1.14

Klein, H. J., & Park, H. M. (2020). Workplace commitments. *In Essentials of job attitudes and other workplace psychological constructs* (pp. 156-183). Routledge.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, *35*, 22199-22213.

Lampinen, A. K., Dasgupta, I., Chan, S. C., Matthewson, K., Tessler, M. H., Creswell, A., McClelland, J. L., Wang, J. X., & Hill, F. (2022). Can language models learn from explanations in context? *Preprint arXiv*. https://doi.org/10.48550/arXiv.2204.02329

Lee, P., Fyffe, S., Son, M., et al. (2023). A paradigm shift from “human writing” to “machine generation” in personality test development: an application of state-of-the-art natural language processing. *Journal of Business and Psychology*, *38*(2), 163–190. https://doi.org/10.1007/s10869-022-09864-6

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology*, *97*(5), 1016–1031. https://doi.org/10.1037/a0027934

Meyer, J. P., & Allen, N. J. (1991). A three-component conceptualization of organizational commitment. *Human Resource Management Review*, *1*(1), 61–89. https://doi.org/10.1016/1053-4822(91)90011-Z

Meyer, J. P., & Maltin, E. R. (2010). Employee commitment and well-being: A critical review, theoretical framework and research agenda. *Journal of Vocational Behavior*, *77*(2), 323–337. https://doi.org/10.1016/j.jvb.2010.04.007

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). Rethinking the role of demonstrations: What makes in-context learning work? *arXiv Preprint*. https://doi.org/10.48550/arXiv.2202.12837

Ng, T. W., & Sorensen, K. L. (2009). Dispositional affectivity and work-related outcomes: A meta-analysis. *Journal of Applied Social Psychology*, *39*(6), 1255–1287. https://doi.org/10.1111/j.1559-1816.2009.00481.x

OpenAI. (2023). *GPT-4 technical report*. Retrieved from https://cdn.openai.com/papers/gpt-4.pdf

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., & Schulman, J. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744. https://doi.org/10.48550/arXiv.2203.02155

Reynolds, L., & McDonell, K. (2021, May). Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–7). https://doi.org/10.1145/3411763.345176

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Schick, T., & Schütze, H. (2021, November). Few-shot text generation with natural language instructions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 390–402). https://doi.org/10.18653/v1/2021.emnlp-main.32

Shin, E. (2021). *Automated item generation by combining the non-template and template-based approaches to generate reading inference test items* (Doctoral dissertation, University of Alberta). https://doi.org/10.7939/r3-75wr-hc80

Shi, J., Lin, H., Wang, L., & Wang, M. (2009). Linking the big five personality constructs to organizational justice. Social Behavior and Personality: an international journal, 37(2), 209-222. https://doi.org/10.2224/sbp.2009.37.2.209

Smith, P. C., Kendall, L. M., & Hulin, C. L. (1969). *The measurement of satisfaction in work and retirement: A strategy for the study of attitudes*. Skokie, IL: Rand-McNally.

Solomon, B. C., Nikolaev, B. N., & Shepherd, D. A. (2022). Does educational attainment promote job satisfaction? The bittersweet trade-offs between job resources, demands, and stress. *Journal of Applied Psychology*, *107*(7), 1227–1241. https://doi.org/10.1037/apl0000904

Spector, P. E. (1997). *Job satisfaction: Application, assessment, causes, and consequences*. SAGE Publications, Inc. https://doi.org/10.4135/9781452231549

Speer, A. B., Perrotta, J., Tenbrink, A. P., Wegmeyer, L. J., Delacruz, A. Y., & Bowker, J. (2022). Turning words into numbers: Assessing work attitudes using natural language processing. *Journal of Applied Psychology*, *108*(6), 1027–1041. https://doi.org/10.1037/apl0001061

Stanton, J. M., Sinar, E. F., Balzer, W. K., Julian, A. L., Thoresen, P., Aziz, S., & Smith, P. C. (2002). Development of a compact measure of job satisfaction: The abridged job descriptive index. *Educational and Psychological Measurement, 62*(1), 173–191. https://doi.org/10.1177/0013164402062001112

Thomas, K. W., & Tymon Jr, W. G. (1982). Necessary properties of relevant research: Lessons from recent criticisms of the organizational sciences. *Academy of Management Review*, *7*(3), 345-352. https://doi.org/10.5465/amr.1982.4285308

Thoresen, C. J., Kaplan, S. A., Barsky, A. P., Warren, C. R., & De Chermont, K. (2003). The affective underpinnings of job perceptions and attitudes: A meta-analytic review and integration. Psychological Bulletin, 129(6), 914–945. https://doi.org/10.1037/0033-2909.129.6.914

Martínez-Tur, V., Moliner, C., & Maniezki, A. (2020). Organizational justice. *Essentials of job attitudes and other workplace psychological constructs*, 45-70.

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

Watson, D., Clark, L. A., & Carey, G. (1988). Positive and negative affectivity and their relation to anxiety and depressive disorders. *Journal of Abnormal Psychology*, *97*(3), 346–353. https://doi.org/10.1037/0021-843X.97.3.346

Watson, D., & Slack, A. K. (1993). General factors of affective temperament and their relation to job satisfaction over time. *Organizational Behavior and Human Decision Processes*, *54*(2), 181–202. https://doi.org/10.1006/obhd.1993.1009

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. *arXiv Preprint*. https://doi.org/10.48550/arXiv.2201.11903

Weiss, D., Dawis, R. V., England, G. W., & Lofquist, L. H. (1967). *Minnesota Satisfaction Questionnaire--Long Form (MSQ)* [Database record]. APA PsycTests. https://doi.org/10.1037/t05540-000

Yao, X., & Zhang, Y. (2010, June). Question generation with minimal recursion semantics. In *Proceedings of QG2010: The Third Workshop on Question Generation* (pp. 68–75).

Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021, July). Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning* (pp. 12697–12706). PMLR. https://doi.org/10.48550/arXiv.2102.09690

APPENDICES

Appendix A

Colquitt et al. (2019) Adapted Content Validity Instructions

Content Validity Instructions

Please read the instructions very carefully. The questions are unique to survey measurement development and require detailed attention.

Research projects in the management field often use survey items to measure work-related concepts, such as work motivation, job satisfaction, and employee stress. When writing survey items, management researchers must take great care to ensure that the items do a good job of measuring the concepts of interest (e.g., that an item meant to measure work motivation really seems to capture that concept well). The goal of this study is to assess survey items used in the management literature.

Your job in this survey is to assess the degree to which each item listed measures a given construct.

On the next few pages, you will see a **bolded construct definition**, followed by several survey items. For each survey item, you will rate how well it measures the **bolded construct definition**. Again, simply rate the degree to which each survey item measures the **bolded** statement on the page. Not all of the items will match the **bolded** statement. Therefore, please pay close attention to each individual item as you decide whether it matches the **bolded** statement.

1	2	3	4	5	6	7
Item does an EXTREMELY BAD job of measuring the **bolded** concept provided below	Item does a VERY BAD job of measuring the **bolded** concept provided below	Item does a SOMEWHAT BAD job of measuring the **bolded** concept provided below	Item does a NEITHER GOOD NOR BAD job of measuring the **bolded** concept provided below	Item does a SOMEWHAT GOOD job of measuring the **bolded** concept provided below	Item does a VERY GOOD job of measuring the **bolded** concept provided below	Item does an EXTREMELY GOOD job of measuring the **bolded** concept provided below

For example, let's say the statement is: Work Motivation: The effort expended in relation to work. Since this statement refers to effort, an item that does a ***good*** job matching this statement might be, "I work hard in my job," because it speaks to a certain effort level at work. An item that also does a good job matching this statement might be, "I often feel lazy at the office," because it also speaks to a certain effort level at work. In contrast, an item that does a ***bad*** job matching this statement might be, "I work in a city," because it has very little to do with the effort level at work. Please note that some of the items on the survey will focus on high levels of a given concept (like the "I work hard" item), whereas others will focus on low levels of a given concept (like the "I often feel lazy" item). Both can capture the concept of expending effort equally well.

Let's Practice! Using the example above, please rate the following three items on how well each does matching our concept, **Work Motivation: The effort expended in relation to work.**

Appendix B

Abridged Job Descriptive Index

Stanton et al. (2002)

Work

1. My job gives me a sense of accomplishment.
2. My job is dull.
3. My job is satisfying
4. My job is uninteresting.
5. My job is challenging.

Pay

1. My job pay is fair.
2. I am underpaid at my job.
3. My job income is adequate for normal expenses.
4. My job is well paid.
5. My job pay is insecure.

Promotion

1. I have a good chance for promotion.
2. I have a dead-end job.
3. My job gives promotions on ability.
4. There are good opportunities for promotion at my job.
5. My job has an unfair promotion policy.

Supervisor

1. My supervisor praises good work.
2. My supervisor is annoying.
3. My supervisor is tactful.
4. My supervisor is bad.
5. My supervisor is up to date.

Coworker

1. My coworkers are helpful.
2. My coworkers are boring.
3. My coworkers are intelligent.
4. My coworkers are lazy.
5. My coworkers are responsible.

Appendix C

Organizational Commitment

Meyer et al. (1993)

Original and revised affective commitment items:

Items denoted with (R) are reversed scored. Items denoted with (RS) are those included in the revised affective commitment measure.

1. I would be very happy to spend the rest of my career with this organization (RS)
2. I enjoy discussing my organization with people outside of it
3. I really feel as if this organization's problems are my own (RS)
4. I think that I could easily become as attached to another organization as I am to this one
5. I do not feel like "part of the family" at my organization (R) (RS)
6. I do not feel "emotionally attached" to this organization (R) (RS)
7. This organization has a great deal of personal meaning for me (RS)
8. I do not feel a strong sense of belonging to my (R) (RS)

Original normative commitment items:

Items denoted with (R) are reversed scored.

1. I think that people these days move from company to company too often
2. I do not believe that a person must always be loyal to his or her organization (R)
3. Jumping from organization to organization does not seem at all unethical to me (R)
4. One of the major reasons I continue to work for this organization is that I believe that loyalty is important and therefore feel a sense of moral obligation to remain
5. If I got another offer for a better job elsewhere I would not feel it was right to leave my organization
6. I was taught to believe in the value of remaining loyal to one organization
7. Things were better in the days when people stayed with one organization for most of their careers
8. I do not think that wanting to be a "company man" or "company woman" is sensible anymore (R)

Continuance

1. I am not afraid of what might happen if I quit my job without having another one lined up.(R)
2. It would be very hard for me to leave my organization right now, even if I wanted to.
3. It wouldn't be too costly for me to leave my organization now.(R)
4. Too much in my life would be disrupted if I decided to leave my organization now.*
5. Right now, staying with my organization is a matter of necessity as much as desire.*
6. I feel that I have very few options to consider leaving this organization.
7. One of the few serious consequences of leaving this organization would be the scarcity of available alternatives

8. One of the major reasons I continue to work for this organization is that leaving would require considerable personal sacrifice—another organization may not match the overall benefits I have here.*

Items denoted with (R) are reversed scored. Items denoted with (RS) are those included in the revised affective commitment measure.

Appendix D

Organizational Justice

Colquitt (2001)

Procedural justice

The following items refer to the procedures used to arrive at your (outcome). To what extent:

1. Have you been able to express your views and feelings during those procedures?
2. Have you had influence over the (outcome) arrived at by those procedures?
3. Have those procedures been applied consistently?
4. Have those procedures been free of bias?
5. Have those procedures been based on accurate information?
6. Have you been able to appeal the (outcome) arrived at by those procedures?
7. Have those procedures upheld ethical and moral standards?

Distributive justice

The following items refer to your (outcome). To what extent:

1. Does your (outcome) reflect the effort you have put into your work?
2. Is your (outcome) appropriate for the work you have completed?
3. Does your (outcome) reflect what you have contributed to the organization?
4. Is your (outcome) justified, given your performance?

Interpersonal justice

The following items refer to (the authority figure who enacted the procedure). To what extent:

1. Has (he/she) treated you in a polite manner?
2. Has (he/she) treated you with dignity?
3. Has (he/she) treated you with respect?
4. Has (he/she) refrained from improper remarks or comments?

Informational justice

The following items refer to (the authority figure who enacted the procedure). To what extent:

1. Has (he/she) been candid in (his/her) communications with you?
2. Has (he/she) explained the procedures thoroughly?
3. Were (his/her) explanations regarding the procedures reasonable?
4. Has (he/she) communicated details in a timely manner?
5. Has (he/she) seemed to tailor (his/her) communications to individuals' specific needs?

Appendix E

PANAS (Negative and Positive Affect)

Watson et al. (1988)

1. Interested
2. Distressed
3. Excited
4. Upset
5. Strong
6. Guilty
7. Scared
8. Hostile
9. Enthusiastic
10. Proud
11. Irritable
12. Alert
13. Ashamed
14. Inspired
15. Nervous
16. Determined
17. Attentive
18. Jittery
19. Active
20. Afraid

Appendix F

Mini-IPIP

Donnellan et al. (2006)

1. I am the life of the party.
2. A Sympathize with others' feelings
3. Get chores done right away.
4. Have frequent mood swings.
5. I Have a vivid imagination.
6. Don't talk a lot. (R)
7. Am not interested in other people's problems. (R)
8. Often forget to put things back in their proper place. (R)
9. Am relaxed most of the time. (R)
10. I Am not interested in abstract ideas. (R)
11. Talk to a lot of different people at parties.
12. A Feel others' emotions.
13. Like order.
14. Get upset easily.
15. I Have difficulty understanding abstract ideas. (R)
16. Keep in the background. (R)
17. Am not really interested in others. (R)
18. Make a mess of things. (R)
19. Seldom feel blue. (R)
20. I Do not have a good imagination. (R)

Appendix G

Burnout

Boswell et al. (2004)

1. I feel emotionally drained from my work
2. I feel burned out from my work
3. I feel exhausted when I think about having to face another day on the job

Appendix H

Job Stress

Solomon et al. (2022)

1. My job is more stressful than I had ever imagined
2. I fear that the amount of stress in my job will make me physically ill.

Appendix I

OCB and CWB

Dalal et al. (2009)

Organizational Citizenship Behavior (OCB)

1. Went out of my way to be a good employee.
2. Was respectful of other people's needs.
3. Displayed loyalty to (organization name).
4. Praised or encouraged someone.
5. Volunteered to do something that was not required.*
6. Showed genuine concern for others.
7. Tried to uphold the values of (organization name).
8. Tried to be considerate to others.

Counterproductive Work Behavior (CWB)

1. Spent time on tasks unrelated to work.
2. Gossiped about people at (organization name).
3. Did not work to the best of my ability.
4. Said or did something that was unpleasant.
5. Did not fully comply with a supervisor's instructions.
6. Behaved in an unfriendly manner.
7. Spoke poorly about (organization name) to others.
8. Talked badly about people behind their back.

Appendix J

Locus of Control

Goldberg (1999)

1. Believe that my success depends on ability rather than luck.
2. Believe that unfortunate events occur because of bad luck.
3. Believe that the world is controlled by a few powerful people.
4. Believe some people are born lucky.
5. Believe in the power of fate.

Appendix K

Phase 2 Demographics

1. Please indicate your age.
2. Please select your gender identity
   A. Man
   B. Woman
   C. Trans-Man
   D. Trans-Woman
   E. Gender Queer
   F. Other
3. Please select the race that best describes you (select all that apply)
   A. African American or Black
   B. American Indian or Alaska Native
   C. Asian
   D. Middle Eastern or North African
   E. Native Hawaiian or Other Pacific Islander
   F. Two or more races
   G. White
   H. Other
4. What is your ethnicity?
   A. Hispanic or Latino/a/x
   B. Not Hispanic or Latino/a/x
5. What is your level of education?
   A. Less than High School
   B. Some High School
   C. High School Diploma or Equivalent
   D. Vocational Training
   E. Some College
   F. Associate Degree
   G. Bachelor's degree
   H. Some Graduate School
   I. Master's Degree
   J. Applied or Professional Doctorate Degree
   K. Doctorate Degree