Journal of Applied Testing Technology, Vol 24(Special Issue), 0-0, 2023

Evaluating the Quality of AI-Generated Items for a Certification Exam

Alan D. Mead[1]* and Chenxuan Zhou[2]

[1]Chief Psychometrician, Certiverse, 4803 N. Milwaukee Avenue, Suite B, Unit 103, Chicago, IL 60630, USA; alan@certiverse.com

[2]Psychometrician, Certiverse, 4803 N. Milwaukee Avenue, Suite B, Unit 103, Chicago, IL 60630, USA; chenxuan@certiverse.com

Abstract

OpenAI's GPT-3 model can write multiple-choice exam items. This paper reviewed the literature on automatic item generation and then described the recent history of OpenAI GPT models and their operation, and then described a methodology for generating items using these models. This study then critically evaluated GPT-3 at the task of writing multiple-choice exam items for a hypothetical psychometrics exam. We also compared two versions of the GPT-3 model (text-davinci-002 and text-davinci-003) on 90 items generated by GPT-3. The vast majority (71% and 90%) of items were judged as useful, but the typical item will require revision to address problems with the stem, key or distractors. The most common error was a violation of the principles of multiple-choice items (e.g., having two correct responses).

Keywords: Item Writing; Artificial Intelligence; Machine Learning; Prompt Engineering

1. Introduction

GPT models from OpenAI have generated considerable interest, particularly since ChatGPT was released on November 30, 2022. As models have grown, each generation of the model (e.g., GPT-2, GPT-3, ChatGPT, GPT-4) has been more powerful in solving Artificial Intelligence (AI) problems (OpenAI, 2023; Brown *et al*., 2020); there is some evidence that large language models improve at a variety of tasks as the models become larger (have more parameters) and require more computation (Kaplan *et al*., 2020). The current paper quantified the quality of exam items written by GPT-3 and compared two versions of the GPT-3 model on this task. We also examine the frequency of errors.

The next part of this introduction summarizes the history of Automatic Item Generation (AIG) before the advent of large language models. Readers eager to learn (in very non-technical terms) how models like GPT-3 operate, and how GPT-3 can be used to generate items may wish to skip ahead, but we see value in a brief review of the history of AIG approaches to provide a context for evaluating the results of recent large language models. The final two sections of this introduction describe our methodology for generating multiple-choice exam items using GPT-3 and then outline our research questions.

Drasgow *et al*. (2006) famously described "strong AIG" and "weak AIG" where weak AIG reflected pragmatic approaches to AIG that have little theory leading to uncontrolled item psychometrics. Weak AIG also encompasses frameworks where a new item's psychometrics can be estimated because of similarity to another item (e.g., the child item is known because the parent item's psychometric characteristics have been established empirically). In contrast, strong AIG would be based on models of behaviour or cognition that would allow more accurate estimating of the psychometric properties (at least difficulty) of the items generated. An example of strong AIG is Embretson's (1998) work

Author for correspondence

generating matrix items to measure fluid intelligence. She demonstrates a method of analyzing the cognitive process of answering these items, which predicts the working memory load of the item, which she finds to strongly predict the item difficulty. In one study (Embretson, 2002) she reports that about 93% of items are useable and seven percent required manual intervention to produce useable items. Almost all strong AIG has focused on cognitive analysis of measures of basic abilities (Bejar, 1993).

The state-of-the-art AIG for professional certification exams and language testing has been characterized by two weak AIG approaches: template-based AIG (Gierl *et al*., 2021) and what might be called "full AIG" (Mead, 2014). In template-based AIG, an item model is authored by a subject-matter expert (SME; Gierl & Lai, 2013). The model includes all text of the item, like any item, but also indicates fields that can be replaced. Some models are created from scratch, but they can also be created by working from an existing item. For example, an item like: "Mary purchased five pencils for $0.25 each. How much did they spend before any taxes?" might become a model like: "<NAME> purchased <QUATITY> <ITEM> for <PRICE> each. How much did they spend before any taxes?" The model also provides some indication of the replacements for the fields. For example, a list of names, a range of quantities, a list of items, and a range (or list) of prices. The model answer might be a formula like " $(Quantity*Price)" and there might be formulas for the distractors. A distinction has been made (Irvine & Kyllonen, 2002) between incidentals like <NAME> and <ITEM> which are surface characteristics of an item that do not affect the item solution and thus are unlikely to change the item's psychometric properties, and radicals like <QUANTITY> and <PRICE> that do affect the solution and can thus potentially be modified to adjust the item's psychometric properties.

Creating such models requires more effort than creating individual items and considerable work has been focused on making it easy for SMEs to author template-based items (Gierl *et al*., 2021). Although there are template-based approaches that generate very dissimilar items for specific domains (e.g., n-layer items; Gierl *et al*., 2021), it is more typical for all items from one model to be considered part of a single family (Sinharay & Johnson, 2005) and therefore to be mutual item enemies that cannot be administered on the same form. Although item models can be carefully written and scrutinized, not all items generated are usable (Gierl *et al*., 2021). In one recent study with two exams, 81% and 93% of the items were accepted. Items were rejected for technical and content flaws and because the item wasn't at grade level (Embretson & Kingston, 2018).

In contrast, the "full AIG" approach uses machine learning models to generate the items entirely from the source material (Gierl *et al*., 2021). A particular concern in this literature has been generating good correct and incorrect response options.

Mead (2014) reported on a project to generate verbal analogies items. Test preparation guides for this item type encourage students to identify the "bridge" (relationship) between words and then apply this bridge relationship to each response option to identify the correct response. Mead assembled a corpus of bridge relationships with associated word pairs. Items were generated by selecting two pairs of words from the same bridge to act as the stem and key. Distractors were generated using pairs with an unrelated bridge. An example item using the bridge "Y operates a(n) X" was:

rocket:astronaut::?

a. lamp:light
b. stick:skating rink
c. jet:pilot
d. demand:supply

In this example item, "rocket:astronaut" and "jet:pilot" were pairs taken from the same bridge and the other pairs were taken from other bridges.

Mead (2014) reported a mean disattenuated correlation of 1.005 with a form composed of items written by human item authors and nearly identical convergent validity coefficients for forms written by humans (0.66) and forms generated by software (0.68). However, the approach had limitations. About 37% of the items generated were unusable, for reasons including over-use of a bridge or a pair (because some bridges had few pairs); ambiguous word pairs (e.g., drum:drum); and inadvertently selecting a distractor pair that operated as an unintended key. In contrast, the main flaw of the manually written items was that they tended to use rare words and were too hard (65% correct, compared to 77% for AIG Items). Also, although Mead's approach generated items from a database of word pairs, his approach could be considered a very complex template because all items have the same basic format and word pairs were drawn from a curated database of bridges and associated word pairs.

Mitkov *et al*. (2006) created multiple-choice items from an online linguistics textbook. Their system searched for sentences about key concepts. In their example, if "syntax" is a key linguistic topic and they found the sentence: "Syntax studies the way words are put together into sentences." then the sentence would be transformed into an item: "Which discipline studies the way words are put together into sentences?" with "syntax" as the correct response. Distractors would be formed by semantically similar words. Their example is that "semantics" and "pragmatics" would form better distractors than semantically more remote terms like "chemistry" or "mathematics." They used a (human) subject-matter expert to edit the items after generation. This editor classified items as "unworthy," which were discarded, or "worthy." Items deemed worthy were further classified into the degree of editing required ("none," "minor," "fair," or "major"). In one study, 43% of 575 items were deemed unworthy and discarded. Of the remaining 57% of the items, the degrees of editing ("none," "minor," "fair," or "major") was 3.5%, 9.0%, 19.3%, and 25.2%, respectively. In a comparison of manually writing items versus using their process, authors were 4.3 times more productive using the automated system (including rejecting unworthy items and editing items).

In a pilot study of the same AIG method in a medical domain (Mitkov *et al*., 2006), 54% of the items were classified as "worthy" and about a third did not require edits to the stem. One issue raised by their subject matter experts was that the critical topic term (e.g., "syntax") is always used as the correct response, whereas some items would have been more effective if the critical topic term was part of the stem and the correct response was a less central topic, presumably because the correct response stood out and made the item trivially easy. ["[T]he anchor of a [multiple-choice test item] produced by RIG always corresponds to a key term. However, the domain experts pointed out several cases in which the key term should stay in the stem and for another less prominent concept to serve as the answer." p. 113]

Cloze items have a missing word that the test-taker must provide or choose (e.g., "Would you like to go __ the movie this afternoon? A. to; B. in; C; at"). They are commonly used in language testing and have been a target for AIG because they are easier to generate by editing existing sentences: select an existing sentence, identify a word to blank (that word becomes the key), and generate distractors. For example, Sumita *et al*. (2005) created language learning Cloze items by sampling sentences from a corpus of "business and travel conversations" and using part-of-speech tagging to determine the sentence's verb, which they then blanked (e.g., "I need to [keep] my head above water."). A thesaurus was consulted to find similar verbs and web searches were conducted to determine whether the terms were used in that context (e.g., "... need to [promise] my head..." does not appear in web searches and therefore would have been accepted as incorrect and therefore a legitimate foil). In one evaluation, 93.5% of 4000 such items were answered correctly by a native English speaker.

Liu *et al*. (2005) reported on a similar project to generate Cloze items for language learners. Items targeting nouns, verbs, adjectives and adverbs were created. The seed sentences were sampled from a corpus of 160,000 sentences with four million words harvested from news websites. The frequency of the target word was discussed as a measure of the difficulty of the resulting item (choosing a sentence with a rare word would create a more difficult item), but item difficulty was not investigated. Distractors were chosen to have similar meanings but not to collocate with important (non-stopword) words in the sentence in the corpus; the acceptable degree of collocation was tuned to match the distractors of a set of existing (manually-created) Cloze items. About 60% of the Cloze items generated were acceptable and had unique answers; Cloze items targeting adverbs were noticeably harder to generate than those targeting nouns, verbs and adjectives. No rigorous analysis of difficulty was presented. Pino *et al*. (2008) also reported a 66% success rate, while Lee and Seneff (2007) reported a 96.3% success rate for the easier task of generating Cloze items targeting prepositions.

To summarize, traditional "full AIG" approaches typically involve considerable effort to create a corpus and machine learning models generated specifically for that domain. Creating and fitting these models is expensive and time-consuming and, in some cases, requires that an existing bank of items has been created (von Davier, 2018). Thus, one of the definitive features of large language models like GPT-3 is that it is pre-trained and thus is comparatively quick and easy to use.

Another characteristic of typical "full AIG" approaches is that only a portion of the items are useable. Although this is true for any item development process, there is a tendency to have a larger percentage of unusable items when more of the item is generated, as opposed to being copied from the corpus.

2. OpenAI Language Models

Because large language models seem "like magic", it is important to understand how these models work.

In February 2019, OpenAI released its first Generative Pre-trained Transformer (GPT) autoregressive model as GPT-2. An autoregressive language model is a type of machine-learning model that predicts the next word in a sequence of words based on the words that have come before it. GPT-2 had 1.5 billion parameters and was trained on a 40GB corpus (called WebText) that included 8 million documents scraped from upvoted Reddit pages. This model was a considerable improvement over the previous models and was fine-tuned to produce medical exam items (von Davier, 2019).

About 17 months later, OpenAI released GPT-3, a model with 175 billion parameters (Brown *et al*., 2020) and trained on a much larger corpus. About 60% of this corpus was based on a filtered version of the Common Crawl dataset from 2016 to 2019. About 22% of the corpus was an expanded version of the WebText database. English-language Wikipedia was 2% of the corpus and "two internet-based books corpora" were the remaining 16% of the corpus. In testing by OpenAI, this general-purpose model often outperformed state-of-the-art ML models that had been tuned for a specific purpose.

To use GPT-3, a set of instructions (called a *prompt*) is submitted to the model, which returns a *response*. The prompt consists of simple English instructions. It is common to perform *prompt engineering*, which typically involves iteratively revising a prompt to achieve better generations (Saravia, 2022). This process will be familiar to anyone who has iteratively tweaked search engine queries to get the best responses but effective GPT-3 prompts tend to be much longer than search engine queries. The response is one or more lines of text composed by the GPT-3 model to "complete" the prompt. The actual mechanism for submission is an API call with a JSON format. The return is also a JSON document.

The English language contains too large a vocabulary to effectively train models on English words. Instead, the model is trained on tokens, which are individual letters, words, or segments of words. The vocabulary size of GPT-3 is 50,257 tokens (Brown *et al.*, 2020); token 0 is "Aardvark", token 1 is "aaron", etc. As a heuristic, OpenAI suggests that 750 words will produce about 1000 tokens. Therefore, a token may be thought of as 0.75 of a word, on average, but many small words are individual tokens while longer words may be broken into two or more tokens. Table 1 shows an example of tokenization.

Grudzien (2022) gives an interesting example of the usefulness of tokenization in helping a model like GPT-3 perform better. He presents the made-up word "overpythonized" and presents evidence that the word is unique (produces no hits from leading search engines). Assuming this word is, indeed, unique, then it cannot have appeared (as a word) in the training data of any model trained before 2022. If GPT models were trained on words, this made-up word would be unique, and the model would know nothing about it. As a result, this word would have to be ignored and any output from a prompt containing this word would not relate to this made-up word. However, this made-up word is tokenized as "over/python/ized", and the model is likely to learn a great deal about the tokens "over," "python," and "ized." As a result, the model could make a reasonable approximation for the meaning of a word like overpythonized. For example, submitting the prompt "define overpythonized:" three times, GPT-3 returned the responses below:

Table 1. An Example of Tokenization

Example	RQ1: How good are the items generated by GPT-3 (using either version of the Davinci models)?
Tokens	R/Q/1/:/ How/ good/ are/ the/ items/ generated/ by/ G/P/T/-/3/ (/using/ either/ version/ of/ the/ Dav/in/ci/ models/)?

Note: Based on: https://platform.openai.com/tokenizer

- "To overpythonize is to take a program written in another language and rewrite it in Python, often to make the code more efficient and easier to understand."
- "To be overpythonized is to be overly obsessed with all things related to the Python programming language. This often includes reading and writing code in Python, attending Python-related events, and networking with other Python enthusiasts."
- "To overpythonize something is to make it overly complex by using advanced Python programming techniques or by adding too many features."

These responses all represent very realistic interpretations of "overpythonize."

As a Markovian model (von Davier, 2019) the GPT models select each next token based on the previous tokens. This is particularly clear when the *temperature* parameter is examined. Using previous text, GPT-3 determines the likelihood of the next token. For example, given part of the well-known mnemonic for the notes corresponding to the lines of the treble clef, "Every good boy deserves" as input, the top five most likely next tokens are: " f" P=0.3842, " favour" P=0.2355, " fruit" P=0.1761, " favour" P=0.1098, and " fun" P=0.0228 (note that the leading space is a part of these tokens). When " f" is chosen, "udge" has a probability of 0.9982 of being chosen next (the OpenAI Playground will display these probabilities; https://platform.openai.com/playground).

Temperature is anthropomorphized as the model being "more creative" in its responses as the temperature approaches its maximum (often 1 or 2). Mathematically, temperature controls the relative importance of the probabilities of the tokens. A higher temperature will make the model more likely to choose tokens with lower probabilities, while a lower temperature will make the model more likely to choose tokens with higher probabilities. When temperature = 0, language models will always choose the most likely next token, and thus always return the same response to a given input (unless other settings are changed). In this example from GPT-3, if the temperature were zero, the model would deterministically return " f" because it has the highest probability. In this study, temperature = 0.7 was used. McCoy *et al.* (2021) warn that "all modifications that increase the novelty of generated text also decrease the quality (p. 5)."

A Markov process is also called a "random walk" and is likely to produce a "word salad" response of tokens that might be related to each other but would not create a coherent whole. The breakthrough that allowed large language models like GPT-3 to create very coherent responses is called "attention" or "self-attention" (Vaswani *et al.*, 2017). At a conceptual level, attention is the aspect of the models that allows them to provide coherent responses to prompts. At a technical level, transformer models contain "attention layers" that are used to help the model better understand the context of a given input. They help the model focus on the important words and phrases in an input sentence and ignore the irrelevant ones. Attention layers also allow the model to dynamically adjust its focus from one part of the input to another. By using attention layers, GPT models can better understand the context of the input and generate more accurate predictions.

The GPT-3 API interface was released as four models of decreasing complexity (Gao, 2021): Davinci (175B parameters), Curie (6.7B parameters), Babbage (1.3B parameters), and Ada (350M parameters). OpenAI recommended that developers engineer prompts for the most capable model (Davinci) and then evaluate whether smaller models performed adequately. The price charged by OpenAI for using Davinci is fifty times more than for using Ada. This difference is widely attributed to the costs associated with running those models. That is, Davinci is a much bigger model that requires more storage, memory, and computing to make a prediction.

After the initial release of GPT-3, the models were improved with Reinforcement Learning from Human Feedback (RLHF), a technique where model outputs are "graded", and the model is tuned to produce a higher "grade" of output (Ouyang et al., 2022). Popular press reports suggest that about 20,000 hours of effort were required to finetune the model (dkirmani, 2022). According to OpenAI (n.d.), "[t]he resulting InstructGPT models are much better at following instructions than [the non-Instruct version of] GPT-3." For example, OpenAI reported that human judges evaluating model output preferred outputs from the InstructGPT Babbage model to outputs from the (pre-Instruct) Davinci model, despite the Babbage model having more than 116x fewer parameters (Lowe & Leikie, 2022). The initial InstructGPT models were identified by "text-<model>-<version>" such as text-davinci-001 (as opposed to the previous models without the "text-" prefix, "<model>-<version>" such as davinci-001).

The GPT models have been called "storytellers" not "truth-tellers" (Kublik & Saboo, 2022) to explain that the model will very confidently assert "facts" that are not true. The ML literature seems to have converged on the term "hallucination" to describe this widespread behaviour (Brown *et al*., 2020). For example, Thorndike's 1904 text "An Introduction to the Theory of Mental and Social Measurements" is generally believed to be the first psychometrics text, but the prompt: "When was the first psychometrics text published?" (with temperature = 0.70) produces answers like these:

- The first psychometrics textbook, "Psychometrics: An Introduction" by Roderick P. McDonald, was published in 1976. [Not true: McDonald authored at least two psychometrics texts long after psychometrics was an established discipline.]
- The first psychometrics text was published in 1904 by psychologist Theodore M. Newcomb, titled "An Introduction to Social Psychology". [Not true: This book was published in 1952 and is not related to psychometrics.]
- The first psychometrics text was published in 1905, written by psychologist Alfred Binet. The book was titled "The Measurement of Intelligence". [Not true: This book was authored by Terman in 1916.]

A tendency towards "hallucination" is hardly surprising when one realizes how these models work, building responses by choosing each successive token based on the current tokens, the training data, and the adaptive "attention." Attention ensures that the responses are related to the topic, but the model has no inbuilt mechanisms to ensure that the response is factual and there is an element of stochastic chance in each response generated by the model (assuming temperature > 0). In other words, GPT-3 "knows" that every good boy deserves fudge (or favour, fruit, favour, fun, etc.) mainly because it has observed that these words form a pattern, but the model has no capability specifically engineered for evaluating the truthfulness of this saying in the way a person would be able to understand this sentence in context.

3. AIG with OpenAI GPT models

3.1 Personality Scales

GPT models have been used to generate exam content. von Davier (2018) reported using deep learning with a recurrent neural network trained on the International Personality Item Pool (IPIP; Goldberg, 1999) to generate Big Five (Goldberg, 1993) personality items like "I worry so much about myself." When a combination of AI-generated and IPIP items was administered to a large sample online, the resulting factor structure recovered

Prompt	Sample Response
Write 10 big-five agreeableness Likert items like these: 1. I make time for others. 2. I am interested in people. 3. I sympathize with others' feelings. 4. I have a soft heart. 5. I take time out for others. 6.	 I am patient with others. 7. I am kind to others. 8. I am considerate of others. 9. I am easy to please. 10. I enjoy being around people.

Figure 1. Example GPT-3 prompt to generate Big Five Agreeableness items.

the intended five factors and loadings for AI-generated items were not noticeably different from IPIP items. This is even more impressive in that the IPIP Items used had been identified as good items for marker scales.

GPT-3 has also been used to generate personality items (Lee *et al*., 2022) by providing examples from the International Personality Item Pool (IPIP; Goldberg, 1999). That is, the prompt was a series of five example items and the scale, and GPT-3 were instructed to write similar items for one of the Big Five dimensions (Goldberg, 1993). For example, the prompt in Figure 1 uses five example Agreeableness items and instructs GPT-3 to write additional items. As in the sample output, GPT-3 generally wrote five more Agreeableness items.

Lee and colleagues generated twenty items for each of the big five personality dimensions and then had subject matter experts select five high-quality items for each of the five domains. Combined with 25 IPIP items, this scale was administered to an online sample. Almost all items (94%) loaded highly on their intended factor, including 100% of the items generated by GPT-3.

These results match our (unpublished) experiences using GPT-3 to write personality items for a psychometrics class exercise. The prompt in Figure 1 is from our work, chosen because it is essentially the same as that used by Lee *et al*. (2022) (Figure 1). Although we did not administer our scales, the items had good face validity and GPT-3 seemed to be roughly as accurate as graduate students in psychology at generating good personality items. We have also used GPT-3 to classify personality statement items into the Big Five using one-shot or few-shot classification (Kublik & Saboo, 2022). We found that GPT-3 was about 80% accurate and had two pathologies: a tendency to produce very plausible classes not trained (e.g., classifying a statement like "I have trouble getting started on chores" into a class like "unconscientious" or "undisciplined" instead of "conscientious"), and having a bias towards classifying items as Neuroticism (e.g., an Extraversion item about shyness is prone to be misclassified). Having GPT-3 classify items into a class like "unconscientious" is acceptable if humans will consume the response. However, this flexibility of response would prevent an automated system that expects one of the big five domains as a class.

3.2 Multiple-Choice Items

As compared to generating Likert statements to measure personality traits, generating multiple-choice exam items is a more difficult AIG task. Such items should align with a chosen exam topic, should have a sensible stem, and should have an appropriate number of response options with a single key. von Davier (2019) fine-tuned GPT-2 using 8GB of open-access medical articles to generate vignettes that could be used for medical exam questions with human editing. He also generated potential distractors from prompts like "What are the most common side effects of statins?" No rigorous evaluation was performed, but the author concluded that the content "was by far not perfect, but they could serve as inspiration for human item writers (p. 12)."

Attali *et al.* (2022) used GPT-3 to generate items for an interactive reading task, including source passages, questions associated with each passage, and correct answers and distractors of each question. In generating the passages, 3-5 examples were provided to GPT-3 for each of the three given genres: news, expositor, and narrative texts, and each example consists of a topic, title, and passage. A total of over 14,000 passages were generated and filtered based on their length, duplicity, and potential offensive content. A sample of 789 passages were retained, and for each of these passages six items were generated: 1 vocabulary-in-context task, 1 text completion task, 2 comprehension tasks, 1 main-idea task, and 1 possible title task. Each passage and question went through at least three content reviews and two fairness reviews, and a final set of 454 passages with their associated questions survived the review process. It is estimated that the review process, across all rounds, took about 15 min per passage. These reading tasks were administered at the end of a practice test on the Duolingo English Test (DET), and nearly 200,000 interactive reading sessions were completed.

These previous works, using either GPT-2 or GPT-3 models, have demonstrated that GPT models can be used to generate usable items for constructing stable and valid personality scales or exam content. While it seems that the application of GPT models in AIG for real, high-stakes, exams and tests is upon us, these previous works also share in common that for the test to work, a filter or review process is necessary. In addition, the usefulness of generated items depends on factors like item type and content domain. The current study attempts to quantify the usability of items generated for a fake, yet realistic, exam.

3.3 Research Questions

In November 2022, text-davinci-003 was released as a replacement for the model Certiverse was using, text-davinci-002. The new version of the largest GPT-3 model was released with the promise of three specific improvements: higher quality writing; better handling of complex instructions (i.e., in a prompt); and better generation of long responses (OpenAI, n.d.).

Generating higher-quality writing would certainly improve item writing, but items are too short to benefit from improvements in long responses. The improvements in long responses could cause items to be longer without being better, which would have been an issue. Finally, (paradoxically) better handling of complex prompt instructions also had the possibility of degrading the performance of the prompts that were engineered specifically for the 002 version of the model (even if prompts engineered for the 003 models might perform better).

Thus, this study was conducted to compare the performance of text-davinci-002 to text-davinci-003. We wanted to evaluate how well these two models worked for writing multiple-choice exam items. We had been using the GPT-3 Davinci model to generate items, but we had completed only one preliminary study. We took the opportunity to address three research questions:

RQ1: How good are the items generated by GPT-3 (using either version of the Davinci models)? We operationalized "good" by judging the quality of the items (see the Method section).

RQ2: Do either of the two versions of the Davinci model produce better items? We operationalized "better items" by judging the quality of the items (see the Method section).

RQ3: What are some common errors in the items that are not ready to use? To address this question, we developed a coding system (see the Method section).

4. Generating items using GPT-3

The Certiverse model for generating items uses a generic prompt with fields for exam-specific information that are filled from the exam blueprint. Table 2 shows values used for exam-level fields and examples of values used for fields that are specific to a blueprint topic or subtopic. Instantiating these values into the prompt template creates a prompt tailored to the specific job role, exam, and blueprint topic and subtopic. However, if two authors generate items for the same subtopic of the blueprint, they should not send identical prompts to GPT-3. Furthermore, item authors should be allowed some degree of input on the prompt to GPT-3. Therefore, the prompt also incorporates a specific item topic that is entered by the item author. See Appendix C for examples.

The exact form of the prompt is considered proprietary, but is an elaboration of "Write a multiple-choice item for a <exam> exam on topic <specific topic>:". For example, using the first specific topic from Appendix C, the prompt would be "Write a multiple-choice item for a Psychometrician exam on topic sample size planning:" Submitting this prompt twice with temperature = 0.70 produced these two items:

Q: What is the primary purpose of sample size planning?
A. To minimize Type I errors
B. To maximize Type II errors
C. To minimize Type II errors
D. To maximize statistical power

Q. Sample size planning is important in research because it:
A. Ensures the validity of the results
B. Minimizes the margin of error
C. Maximizes the statistical power
D. All of the above

The fact that GPT-3 produces multiple-choice items in this format is, in our experience, unique to the OpenAI GPT models. Other models that we evaluated before 2022 did not "know" the correct format of a multiple-choice item. It is impressive that GPT-3 can produce such items with no specialized training in item writing, formats, or psychometrics. Furthermore, these two items seem to be at a comprehension or application level, rather than being trivial recall questions.

Table 2. Fields used to Generate Psychometrician Items

Prompt Field	Values
Actual values	
Exam Description	Psychometrics exam
Job/Role title	Psychometrician
Job/role description	responsible for the technical quality of educational assessment and analysis activities. Duties include job analysis, item-writer training, creating and maintaining exam blueprints, item writing and review, pretesting, setting performance standards, cut scores, item and test analysis, equating, item bank design and analysis, and other psychometric activities. The individual will lead technical operational and research projects, ensure quality control of deliverables, perform and monitor statistical analyses, conduct research and special analyses, and contribute to the development of data interpretation materials and publications. This position requires an understanding of the testing industry, mathematical and statistical theory.
Example values	
Topic	Analysis
Subtopic	Planning analyses, Knowledge of: analytic methods and feasibility; sample size requirements and statistical power; data collection design and methods. Ability to document research questions and hypotheses.
Specific question topic	sample size planning

On the other hand, these items are not high-quality polished items measuring very high-level cognitive skills. The first item has two correct answers (C and D). This item could be fixed by rewriting "C" or "D" to be incorrect. The second item uses "All of the above," which is generally not encouraged (Haladyna & Downing, 1989), and has no key because the answer "A" is not technically correct but both "B" and "C" are correct making "D" incorrect. This item could be fixed by rewriting "A" to be correct (then "D" would be the key) or (preferably) by rewriting "D" and either "B" or "C" to be incorrect distractors. [A reviewer correctly pointed out that the GPT model could make these corrections; our point is that there are specific flaws that appear in these items.]

Although it is possible that additional, domain-specific fine-tuning might produce better items, particularly if there is a domain-specific vocabulary (von Davier, 2019), generating items with GPT-3 has several advantages. First, anyone who can write simple English instructions can write a GPT-3 prompt. Second, no special preparation is needed; no corpora need to be gathered, and no models need to be fit. Compared to methods that use banks of items as the corpus, GPT-3 can be used for new exams or new topics on existing exams.

5. Method

5.1 Development of a Psychometrician Exam Blueprint

Our method of generating items (see below) requires various values that would be taken from an exam and exam blueprint. Therefore, we decided to build a realistic exam blueprint. See Appendix A for details.

5.2 Generating Items

In prior work, we performed "prompt engineering" (Saravia, 2022; Kublik & Saboo, 2022), which is a process similar to iteratively tweaking search queries to extract the correct information from a search. We started with a simple prompt, and modified it, mainly by adding additional instructions, and evaluated output responses (multiple-choice items) iteratively to attempt to produce better results. Certiverse considers the exact form of the prompt to be proprietary, but the prompt is a more elaborate form of the simple prompts discussed in the previous section. The prompt includes fixed instructions to write an item as well as information about the exam, the role, the specific blueprint topic, and the specific item topic (Table 2).

Using this prompt, two parallel sets of 45 items were generated for topics on a hypothetical psychometrician blueprint. The prompts for each item were generated to address the specific blueprint topic and specific item topics (see Appendix C). For each item, an identical prompt was submitted to the text-davinci-002 and text-davinci-003 engines via API calls. Appendix C shows the list of topics used in the prompt for generating the 45 pairs of items. Appendix E shows the 90 items (45

Table 3. Examples of Items rated in Each Category

Rating	Example Item
1	6. You are planning to analyze the data collected from a study. Which of the following methods would be the most appropriate for this type of data? A. Mean B. Median C. Mode **D. Standard deviation**
2	1. You are planning to conduct a study to compare the effectiveness of two different teaching methods. Which of the following would be the best sample size to use in order to have enough power to detect a small difference between the two methods? A. N = 2 B. N = 10 C. N = 50 **D. N = 100**
3	7. You are planning to conduct a study to compare the mean number of hours that full-time employees work in different countries. Which of the following statements is true? A. Non-parametric hypothesis tests should be used when the data are not normally distributed. **B. Non-parametric hypothesis tests should be used when the population variance is unknown.** C. Non-parametric hypothesis tests should be used when the sample size is small. D. Non-parametric hypothesis tests should be used when the data are ordinal.
4	14. A researcher is interested in knowing whether there is a difference in the mean scores of two different groups on a particular measure. She decides to use an independent sample t-test to analyze her data. Which of the following is NOT an assumption of the independent sample t-test? A) The two groups have equal variances **B) The two groups have equal means** C) The two groups are independent of each other D) The dependent variable is normally distributed

Note: Ratings were defined as 1 = Useless; 2 = Needs considerable editing; 3 = Needs minor editing; 4 = Ready as-is; Bolded response was reported as the correct response. Example items for ratings 1 to 4 are all from the 002 models, items: 6, 1, 7, 14. Item numbers were added for clarity.

using each engine) generated and samples are shown in Table 3.

The generated items were evaluated by the two authors in terms of their usability. The usability ratings were in Likert format with 4 points where 1 = useless, 2 = needs much work, 3 = needs minor edits, and 4 = acceptable as is. The usability ratings of the two raters correlated at 0.518 across all items (0.472 for text-davinci-002 items and 0.596 for text-davinci-003 items). Appendix D shows the individual ratings for each item. The average rating of the two raters was then used to compare the performance of the two engines.

To code errors, we started a list of the errors we saw and grouped these into five categories. This included one category that we have seen in previous work (low cognitive complexity), but which was not seen in these items. The first category included several specific issues that made the item invalid as a multiple-choice item, such as not having the correct number of responses, having no key, having two keys, etc. The other errors were having trivial content, being off-topic, and not having sufficient information in the stem to justify the key (Table 4 for examples).

Table 4. Examples of Items with Common Problems

Error Type	Example Item
Failed to meet the multiple-choice item requirements, including no correct response, more than one correct response, missing the question, incorrect key, and wrong number of generated responses.	3. When setting up a study, a Psychometrician needs to understand statistical power and type I error in order to: A. Accurately interpret results. **B. Set up the sample size requirements.** C. Estimate the analysis methods and feasibility. D. Estimate data collection design and methods.
Items with low Bloom's taxonomy levels (e.g., "What is the definition of XXX?")	N/A
Trivial or insignificant content	21. Which of the following best reflects the number of items needed on a typical IT certification exam? A. Between 10 and 20 B. Between 20 and 30 C. Between 30 and 40 **D. Over 40**
Off-topic or content that seems to be related to the topic but making no sense.	16. You are conducting a research project which requires analyzing the data collected. What type of analysis should you use? a. Standardized Test Theory b. Factor Analysis c. Logistic Regression **d. Linear Regression**
Not enough information given in stem	17. Which of the following is the best estimate of the number of cases needed for a job task analysis survey? A. 10-20 **B. 20-30** C. 30-40 D. 40-50

Note: All example items were generated by text-davinci-003

6. Results

RQ1. Research question 1 was “How good are the items generated by GPT-3 (using either version of the Davinci models)?” To address RQ1, we calculated the mean, median and modal ratings for all items. The mean, median and mode of ratings were all about 2, indicating that the average item will need considerable editing (Table 5). If items with a rating of 1.0 (only) are considered useless, then 88.9% of the items would be considered useful. If items with a rating of 1.5 were considered useless, then 71.1% of the items would be considered useful. Thus, most of the items would be considered useful. (However, human subject-matter experts might reasonably be expected to produce close to 100% useful items.) Very few items were rated a 4, indicating that all items will need a review by a SME; most items will require some editing. [A reviewer points out that SME review may well be needed (e.g., for bias, appropriateness to a topic, etc.) even if all items were rated a 4.]

RQ2. Research question 2 was, “Do either of the two versions of the Davinci model produce better items?” To address RQ2, we computed the mean, median and modal rating separately for the two models (See Table 5). The mean and median were higher for the 003 version of the model however the differences were small (Cohen’s $d = 0.15$) and both models produced similar items. An independent sample t-test was not significant, $t(88) = 1.44$, $p > 0.05$. Thus, although the results for the 003 model were slightly better, this was not a statistically significant difference in this sample size and was practically small. Two samples of $N = 50$ have 80% power for a medium effect size; the sample needed to detect a $d = 0.15$ difference with 80% power would be 550.

While it is possible that low inter-rater reliability (r_{12}=0.518) contributed to the null results, it is worth noting that the mean ratings of each rater were close to 2. Thus, although the agreement between the raters was imperfect, both agreed that the typical item from both models required substantial edits.

RQ3. Research question 3 was, “What are some common errors in the items that are not ready to use?” To address RQ3, we examined the items. Table 6 summarizes the kinds of errors we observed. The two groups did not differ significantly in the types of errors observed, $X2(df=4, N=90) = 3.62$, $p > 0.45$. Table 4 presents examples.

Table 5. Frequency Distribution and Descriptive Statistics of the Usability Ratings

Mean Ratings	text-davinci-002		text-davinci-003	
	frequency	**%**	**frequency**	**%**
1.0	5	11.10%	4	8.90%
1.5	8	17.80%	5	11.10%
2.0	12	26.70%	9	20.00%
2.5	7	15.60%	14	31.10%
3.0	9	20.00%	10	22.20%
3.5	3	6.70%	3	6.70%
4.0	1	2.20%	0	0.00%
Sum	**45**	**100.00%**	**45**	**100.00%**
Mean	2.22		2.33	
SD	0.76		0.67	
Median	2		2.5	

Table 6. Common Problems

Error type	text-davinci-002		text-davinci-003	
	frequency	**%**	**frequency**	**%**
Failed to meet the multiple-choice item requirements, including:	21	46.7%	22	48.9%
no correct response	6	13.3%	9	20.0%
more than one correct response	8	17.8%	10	22.2%
incorrect key	5	11.1%	3	6.7%
missing the question	1	2.2%	0	0.0%
wrong number of responses generated	1	2.2%	0	0.0%
Items with low Bloom's taxonomy levels (e.g., "what is the definition of XXX?")	0	0.0%	0	0.0%
Trivial or insignificant content	1	2.2%	4	8.9%
Off-topic or content that seems to be related to the topic but makes no sense.	11	24.4%	5	11.1%
Not enough information given in stem	6	13.3%	6	13.3%
items with error identified	**39**	**86.7%**	**37**	**82.2%**
items generated	**45**		45	

Our inter-rater reliability was modest. The correlation of our ratings was 0.52, which is an estimate of our individual rating reliability. The average of our two ratings should have a reliability of 0.68. We believe that our modest reliability is because of the difficulty of evaluating the usefulness of a batch of flawed items. As an extreme example, consider model 002 item 19, which had ratings 1 and 4 (useless and ready to go). How can we explain such extreme differences in evaluations? The item was based on the specific topic "how often should a JTA be repeated," which could easily form a simple stem. And, indeed, the item was generated without a stem but just four response options for that question:

A. Once every five years.
B. As needed, but at least once every five years.
C. As needed, but at least once every three years.
D. As needed, but at least once every two years.

The rating of 1 reflected that the item was generated without a stem, and the rating of 4 assumed that the stem was "How often should a JTA be repeated?" Although this specific issue was not common, it reflects part of the difficulty of evaluating these items. The other difficulty was that the specific value below 4 depends on the rater's creativity in imagining a fix for the item.

Another issue with this summarization is that many items were flawed in ways that could be remedied in various ways. Thus the "failed to meet the multiple-choice item requirements" is the most common category, affecting almost half the items, because it includes several flaws, including problems with the stem, key/keys, and distractors. Item 3 by text-davinci-003 is an example. All the responses could be argued to be correct answers to the stem and statistical power and Type I errors are influenced by different factors. Perhaps the best way to fix this item is to ask about why sample size requirements matter in the stem and include "understanding statistical power" as a key.

The second most common issue, affecting 18% of the items, was content that was off-topic or nonsensical. Item 16 by text-davinci-003 is an example that looks on-topic but doesn't provide the elements needed to measure a specific objective. The scenario fails to motivate the key or provide the information needed to distinguish the key from the distractors.

The remaining errors were less common. We observed no items about mere recall in this study, although we have observed that in other (unpublished) work, so we report this category with zero observations. Item 21 by text-davinci-003 is an example of a question asking about trivial content. Item 17 by text-davinci-003 is an example of an item that does not provide enough information in the stem to verify that the key is correct or to distinguish the key from the distractors.

7. Discussion

This research evaluated the items written by GPT-3 and compared versions of the Davinci model (text-davinci-002 and text-davinci-003). To evaluate the models, we created a plausible blueprint for a domain in which we were subject matter experts (psychometrics). We then crafted a series of specific item topics and generated items using the same prompt to address those specific item topics. This procedure matches how we expect GPT-3 to be used to generate items in practice.

Our results suggest that most items (between 89% and 71%) generated by GPT-3 would be useable, but most of those items would require editing, with most of them requiring substantial editing to address unclear stems, no correct answers, or multiple correct answers. The two versions of the Davinci model produced similar results, although "text-davinci-003" might be slightly better. We doubt that SMEs would notice this difference in practice. By far the most common error we observed was a violation of the basic standards of multiple-choice items (no key, two keys, stem unrelated to responses, etc.).

Is this a pessimistic take on the usefulness of GPT-3? When Certiverse integrated GPT-3 into our item writing process, the SME chose to use or skip AI assistance. Item writers who decline to use GPT-3 go into the normal item writing wizard and see blank fields. Those who choose to use AI assistance enter a specific question topic and then are presented with three AI-generated items. They can choose to accept any one of these or to reject all three and see several more (up to 10 items). We evaluated the "first" item generated by GPT-3, which may not generalize well to a single item that an SME chooses from up to 10 candidate items generated by GPT-3. Generating multiple versions is an option in GPT-3 and it is currently (as of this writing) the default with Bard [an LLM provided by Google].

Thus, in practice, maybe a simple solution to significantly improve the utility of a fallible model is to simply sample more than a single item (e.g., three items). To test this, we performed a simulation. We took our distribution of evaluations and, for each item in our sample, we randomly sampled three evaluations and selected the maximum evaluation. This simulates drawing three items from the same population as our sample of 90 items and selecting the top item. We repeated this simulation five times. The mean evaluation ranged from 2.86 to 2.94 with an overall mean of 2.89. The median was 3 for all five replications. This suggests that, in practice, better items (items requiring only minor edits) can be selected simply by sampling more items – three items may be sufficient – and having an SME select the best item of these three.

Although we did not rate our item in terms of cognitive complexity, we noted that the items generated by GPT-3 were also not disproportionately of low cognitive complexity (the items are available in Appendix E). Based on a few casual trials, we do not believe GPT-3 can specifically respond meaningfully to instructions to write at a particular level of Bloom's taxonomy. However, the specific wording of the prompt might encourage or discourage a response in various levels of Bloom's taxonomy. Although we did not write our item topics (see Appendix C) with Bloom's level in mind, we feel that none of our specific item topics encouraged GPT-3 to write trivial items. A larger sample of item topics would be necessary to investigate how the wording of the prompt influences cognitive complexity.

Will these results generalize to domains other than psychometrics?

One of the significant advantages of GPT-3 is that it is a generalist model. That is, rather than having been trained to perform one task, it is trained on a wide body of knowledge. However, how well can it specialize in the myriad domains in which SMEs write exam items? It is impossible to empirically analyze the generalizability of these results to other domains from this sample, but psychometrics is a fairly technical domain. While psychometric content is accessible on the Internet, it must represent a tiny proportion of the content used to train the model. This analysis leads us to expect that many other domains would have similar results, but this will be conjecture until additional studies of this nature have been conducted.

For domains which have comparably much less or no Internet-accessible content, the results would, presumably, be much worse. For domains that are covered widely, the wide coverage may improve the items, but if the quality of that content is low or variable, then it is possible that the domain would have worse performance. One other problem could occur with other domains: A very specific domain (e.g., Illinois privacy law) might have worse results because of wide coverage of the topic (e.g., privacy laws in the US, Europe, etc.) that confuse (e.g., a privacy item intended for Illinois might reference privacy law from another jurisdiction). These problems may also be subtle, requiring more careful analysis and checking by subject matter experts.

Given the literature we reviewed, our experience using GPT-3 to develop Likert statements for personality scales, and our experience using GPT-3 to generate multiple-choice items, we fully expect that this model can assist in generating content for a wide variety of types of instruments and exam domains. More advanced models released after GPT-3 will, presumably, perform even better because research on other benchmarks suggests that newer models outperform prior smaller (or less capable) models (e.g., OpenAI, 2023).

Will AI replace SMEs?

Models like GPT-3 will not replace SMEs for two reasons. First, our AI-generated items frequently violated the basic principles of writing multiple-choice items. One might be tempted to conclude that a generative model that assembles items based on choosing tokens randomly based on previous tokens creates many poorly formed items. However, human authors also violate basic principles of writing multiple-choice items and it is notable the many rules for which we did not observe violations: although we know these models have some potential for biased or offensive language, we observed none. The response options are remarkably well-formed, being fairly parallel, of similar lengths, and usually being quite plausible.

It also seems unlikely to us (and to at least one reviewer) that human review of items would remain an important safeguard, ensuring that items are sound, unbiased, on topic, etc.

The point is not whether the GPT-3 model is a suitable replacement for human authors (it is not) but whether an author is more productive by editing these items. Our opinion is that AI-assisted item writing is more efficient, but the degree of efficiency will have to wait for future research. But in the future, when models have improved and empirical evidence suggests that the engineering challenges of item-writing are solved, very few SMEs may be needed to generate exam content.

Can Our Approach to Prompting be Improved?

This research represents an early attempt to use large language models to write items and the prompting methodology can likely be improved.

Our prompt used a zero-shot approach. That is, the prompt instructed the model to write an item without including any examples of good items in their prompts. This was partially a limitation of our artificial context (we had no existing psychometrician items to use). It would have been interesting to see the results of a one-shot or few-shot approach in which one or a few examples are provided as part of the prompt (see Figure 1). One reviewer noted that this "levels the playing field" relative to human item writers (who are typically provided with item examples as part of their training). An even stronger approach would be to fine-tune (Ouyang *et al.*, 2022) the language model on effective items. Future research should examine the relative benefits of these approaches.

Since we started this research, chat and the concept of prompt chaining have become more commonly used. Chat uses a context of back-and-forth responses (between the user and the model) which often produces more effective responses. Prompt chaining refers to the practice of using a sequence of prompts or questions to guide the model's responses in a coherent and contextually relevant manner. Instead of providing a single prompt or question to the model, users provide a series of prompts sequentially to build upon previous responses and maintain a consistent conversation or line of thought.

While the effect of chat should be evaluated, prompt chaining suggests asking the model to evaluate the item. For example, after generating an item, the model would be asked to classify whether the key was, in fact, the correct answer or not. And, whether each distractor was incorrect. At the time we started this research, it seemed tautological that the model would validate its generations, but the subsequent application of prompt chaining strongly suggests that requiring the model to police itself has the potential to substantially improve the quality of the items generated.

Our approach represents a starting point for understanding how well models like GPT-3 can generate items.

Applicability of These Results to Later models (GPT-3.5, GPT-4)

The pace of innovation and investment in large language models has become intense and GPT-3 is now an obsolete technology. GPT-3.5 (the engine underlying ChatGPT) introduced a chat format (apparently) without fundamentally changing the underlying GPT-language 3 model. We have (unreported) evidence suggesting that GPT-3 003 and GPT-3.5 have identical performance at this task of writing items. It is to be expected that future models will be somewhat better performance. GPT-4, for example, is reported to hallucinate less than GPT-3.5 (OpenAI, 2023). However, informal testing with GPT-4 does not suggest that it produces perfect items. Although we expect that models like GPT-4 and future models perform better, we speculate that the types of issues (no correct answer, two correct answers, etc.) will continue to be an issue until models succeed in preventing hallucinations.

8. Limitations and Future Directions

This is a single study of a single domain. Replication of this study, or actual use, in a variety of domains, will be necessary to evaluate how well these results generalize to other domains.

If the goal of AIG is to produce items without human intervention, that goal seems to remain unmet. As impressive as GPT-3 is, it still hallucinates (creates false information; Kublik & Saboo, 2022; OpenAI, 2023) and makes errors. ChatGPT (i.e., GPT-3.5) and GPT-4 have been released since this research was conducted. GPT-4 is reported to be more accurate (OpenAI, 2023).

It is also worth noting that the AIG methods used in this paper are “weak AIG” in the Drasgow *et al.* (2006) framework because the psychometric properties of the items are unknown. It is not clear how prompts could be engineered to remedy this. Large language models do not have an inbuilt sense of easy, medium, or difficult items. Human subject matter experts cannot routinely agree perfectly about the difficulty of an item and deep domain knowledge, specific to the exam content, would probably be needed to craft prompts that accurately target difficulty levels. Thus, pre-trained models may never be able to predict difficulty. In our experience, high IRT discrimination values result from clearly written items that successfully engage a specific, meaningful learning objective with effective distractors. It is even harder to imagine how to craft a prompt to ensure consistently good item-total correlations. Fine-tuning GPT models on large item pools might increase the effectiveness of the model in generating items with superior discrimination parameters, and future research should investigate this approach.

We wanted to evaluate the difficulty level of items written with GPT-3; for example, does GPT-3 tend to write easy or hard items? And are the difficulty levels of items similar or varied? We could address these issues for the edited items, but it was impossible to meaningfully address these issues for a group of items where the typical quality needed substantial revision. Having no correct answer or having more than one would make the item extremely difficult, but what matters is the difficulty of the finished product. Assuming that items have a substantial human review, a program can control the difficulty of AI-assisted items in the same way that they do today with items written without AI assistance (instructing the item-writers to target a particular level or levels).

Can we reach any conclusions about item quality without administering items and evaluating the psychometric properties empirically? While we agree that empirical estimates of the item's psychometric properties should be evaluated in future research, we believe that exam programs do a good job of screening out defective items. The median percentage of items developed on the Certiverse platform that survive beta testing is 92.5%. We expect comparable results for items written with AI assistance. Thus, we can indeed reach useful conclusions from an evaluation of the items themselves.

9. Conclusion

GPT-3 easily generates items for topics like psychometrics using simple English prompts. Our analysis suggests that most items would be useable but would require human editing. Models will need to improve substantially upon GPT-3 before they will substantially or entirely replace human subject matter experts. We anticipate that models like GPT-3 can assist human subject matter experts in writing items more efficiently and assisting SME item authors seems like a more realistic near-term use of these models in exam content development.

10. Acknowledgements

We thank Kirk Becker, Jialin Huang, Brendan Neuman, and Nathan Thompson for their insightful comments on drafts of this paper. Small parts of this paper were written by the GPT-3 and Bard AI models, however, the authors vouch for, and take full responsibility for, the accuracy of every single token of this paper.

11. References

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, *5*, Article 903077. https://doi.org/10.3389/frai.2022.903077

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323–359). Hillsdale, NJ: Lawrence Erlbaum.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., …, & Amodei, D. (2020). *Language Models are Few-Shot Learners*. ArXiv. https://doi.org/10.48550/arXiv.2005.14165

dkirmani. (2022). *OpenAI's alignment plans* [Blog]. LESSWRONG. https://www.lesswrong.com/posts/28sEs97ehEo8WZYb8/openai-s-alignment-plans

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. In S. H. Irvine, & P. C. Kyllonen (Eds.), *Educational measurement* (4th ed., pp. 471-516). Washington, DC: American Council on Education.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*, 300-396.

Embretson, S. E. (2002). Generating abstract reasoning items with cognitive theory. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (p. 219-250). Mahwah, NJ: Erlbaum.

Embretson, S. E., & Kingston, N. M. (2018). Automatic item generation: A more efficient process for developing mathematics achievement items? *Journal of Educational Measurement, 55*, 112–131.

Gao, L. (2021). *On the sizes of OpenAI API Models* [Blog]. EleutherAI. https://blog.eleuther.ai/gpt3-model-sizes/

Gierl, M. J., & Lai, H. (2013). Using automated processes to generate test items. *Educational Measurement: Issues and Practice, 32*, 36-50. https://doi.org/10.1111/emip.12018

Gierl, M. J., Lai, H., & Tanygin, V. (2021). *Advanced methods in automatic item generation.* New York, NY: Routledge.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality Psychology in Europe*, Vol. 7 (pp. 7-28). Tilburg, The Netherlands: Tilburg University Press.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*, 26-34.

Grudzien, P. (2022). *GPT-3 tokens explained - what they are and how they work* [Blog]. QuickChat. https://blog.quickchat.ai/post/tokens-entropy-question/

Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*, 37-50. https://doi.org/10.1207/s15324818ame0201_3

Irvine, S. H., & Kyllonen, P. C. (2002). *Item generation for test development.* Hillsdale, NJ: Lawrence Erlbaum.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D (2020). *Scaling laws for neural language models.* https://arxiv.org/abs/2001.08361

Kublik, S., & Saboo, S. (2022). *GPT-3: Building innovative NLP products using large language models.* Sebastopol, CA: O'Reilly.

Lee, P., Fyffe, S., Son, M., Jia, A., & Yao, Z. (2022). 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,* 163-190. https://doi.org/10.1007/s10869-022-09864-6

Lee, J., & Seneff, S. (2007). Automatic generation of cloze items for prepositions. In *8th Annual Conference of the International Speech Communication Association*. http://dx.doi.org/10.21437/Interspeech.2007-592

Liu, C.-L., Wang, C.-H., Gao, Z.-M., & Huang, S.-M. (2005). Applications of lexical information for algorithmically composing multiple-choice cloze items. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP.* http://dx.doi.org/10.3115/1609829.1609830

Lowe, R., & Leike, J. (2022). *Aligning language models to follow instructions* [Blog]. OpenAI. https://openai.com/research/instruction-following

McCoy, R. T., Smolensky, P., Linzen, T., Gao, J., & Celikyilmaz, A. (2021). *How much do language models copy from their training data? Evaluating linguistic novelty in text generation using RAVEN.* https://doi.org/10.48550/arXiv.2111.09509

Mead, A. (2022). *Next-Generation JTA* [Webinar]. Certiverse. https://certiverse.com

Mead, A. (2014). *Automatic generation of verbal analogy items.* Unpublished paper.

Mitkov, R., Ha, L. A., & Karamanis, N. (2006). A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering, 12*, 177-194.

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

OpenAI. (n.d.). *How do text-davinci-002 and text-davinci-003 differ?* [Blog.] OpenAI. https://web.archive.org/web/20230314223727/https://help.openai.com/en/articles/6779149-how-do-text-davinci-002-and-text-davinci-003-differ

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., ... & Lowe, R. (2022) *Training language models to follow instructions with human feedback.* https://doi.org/10.48550/arXiv.2203.02155

Pino, J., Heilman, M., & Eskenazi, M. (2008). A selection strategy to improve cloze question quality. In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*. https://www.philippe-fournier-viger.com/ill-defined/w-its08-Pino.pdf

Saravia, E. (2022). *Prompt Engineering Guide.* Github. https://github.com/dair-ai/Prompt-Engineering-Guide

Sinharay, S., & Johnson, M. (2005). Analysis of data from an admissions test with item models. *ETS Research Report RR-05-06*. http://dx.doi.org/10.1002/j.2333-8504.2005.tb01983.x

Sumita, E., Sugaya, F., & Yamamoto, S. (2005). Measuring non-native speakers' proficiency of English by using a test with automatically-generated fill-in-the-blank questions. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP. Association for Computational Linguistics*. http://dl.acm.org/citation.cfm?id=1609829.1609839

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need.* ArXiv. https://doi.org/10.48550/arXiv.1706.03762

von Davier, M. (2018). Automated item generation with recurrent neural networks. *Psychometrika, 83*, 847–857.

von Davier, M. (2019, August 26). *Training Optimus Prime, M.D.: Generating Medical Certification Items by Fine-Tuning OpenAI's gpt2 Transformer Model.* https://doi.org/10.48550/arXiv.1908.08594

Appendix A. Development of a Psychometrician Exam Blueprint

In order to evaluate the quality of items generated by GPT-3, it was necessary for the authors to be subject matter experts. Therefore, we chose to model our item development on a hypothetical "Psychometrician Exam."

We also needed a blueprint for this exam, to evaluate the appropriateness of an item to the intended topic and because our method of generating items requires various values that would be taken from an exam blueprint. Therefore, we decided to build an exam blueprint.

We began with task elicitation. Tasks performed by a psychometrician were collected from O*NET occupations Statisticians (15-2041.00) and Data Scientists (15-2051.00). After removing duplicate and non-crucial entries, 59 tasks were kept, and one or more broad domains were assigned by the authors to each of the tasks. The table below shows the tasks and their associated domains. Eight broad domains were assigned across tasks, which then became 7 root-level topics (two domains were grouped into one) of the exam blueprint.

Rather than conduct a job analysis survey, we exploited a known issue with job analyses: The number of tasks associated with each domain is correlated > 0.90 with the weight derived from a typical job analysis survey (Mead, 2022). This occurs because ratings of task statements on a 5-point scale are generally in the range 3.0 to 5.0, while the number of tasks associated with a domain can vary dramatically. As a result, the initial topic weights can be accurately estimated from the counts of statements per domain. Subsequently, the root-level topics were then broken into 2 to 6 sub-topics and assigned a consensus weight by the study authors. Appendix B shows the final exam blueprint with 7 root level topics and 23 sub-topics.

Again, the purpose of developing this blueprint was to satisfy the requirements of our model for generating items using GPT-3 and to provide a framework for evaluating the items. Any blueprint could have been used. We chose a methodology designed to mimic the way actual blueprints are created in practice in order to maximize generalizability of our results.

#	Task	Domains
1	Analyze data to identify trends or relationships among variables.	Analysis
2	Apply mathematical principles or statistical approaches to solve problems in scientific or applied fields.	Analysis; Psychometrics
3	Design research studies to obtain scientific information.	Research
4	Design software applications.	Software
5	Determine appropriate methods for data analysis.	Analysis
6	Evaluate data quality.	Data management; Psychometrics
7	Evaluate project designs to determine adequacy or feasibility.	Research
8	Evaluate technical data to determine effect on designs or plans.	Research
9	Review items for adherence to quality standards	Psychometrics; Experience
10	Install and maintain statistical analysis software.	Software; Analysis
11	Prepare analytical reports.	Documentation
12	Prepare data for analysis.	Data management; Psychometrics
13	Prepare graphics or other visual representations of information.	Documentation
14	Present research results to others.	Documentation

#	Task	Domains
15	Write computer programming code.	Software
16	Standardize, find efficiencies, and develop infrastructure in support of exam development and maintenance activities	Experience
17	Support and review the products/deliverables of test development, maintenance, and security processes	Experience
18	Clearly communicate the results of psychometric analyses	Documentation
19	Facilitate workshops	Consultation
20	Conduct or assist with validation studies	Research; Psychometrics
21	Knowledge of scaling and equating	Psychometrics
22	Design scaling and equating approaches	Psychometrics
23	Evaluate reasonableness and troubleshoot unusual equating results	Analysis; Psychometrics
24	Recommend appropriate analysis approaches	Analysis; Consultation
25	Construct tests and evaluate tests built by others	Research; Psychometrics
26	Develop psychometric targets for test construction and create test construction specifications	Psychometrics; Experience
27	Plan analysis review meetings and prepare materials	Documentation; Analysis
28	Plan, design, and create materials for standard setting meetings	Consultation; Psychometrics
29	Articulate broad public opinion and national policy concerns about testing	Experience
30	Knowledge of statistical software	Software
31	Knowledge of use of IRT software in a research or operational environment	Psychometrics; Software
32	Knowledge of implementing different IRT models	Psychometrics
33	Lead job analysis projects	Research
34	Knowledge of Data management	Data management
35	Knowledge of Scoring	Psychometrics
36	Knowledge of Test and item analysis	Psychometrics
37	Knowledge of Equating, linking, and scaling of tests	Psychometrics
38	Knowledge of Item bank maintenance	Experience; Psychometrics; Software
39	Knowledge of Quality control	Data management
40	Knowledge of Technical documentation	Documentation
41	Ensures quality control of deliverables	Data management; Experience
42	Performs and monitors statistical analyses	Analysis
43	Conducts research and special analyses	Research
44	Assists in planning, coordinating, and conducting statistical work	Analysis; Psychometrics
45	Represents organization's position on technical issues	Documentation

#	Task	Domains
46	Applies knowledge of statistical procedures, psychometric methods, and statistical programming	Analysis; Psychometrics; Software
47	Consult with clients to identify needs	Consultation
48	Develop and maintain exam infrastructure	Software; Experience
49	Review and develop test development, maintenance, and security processes	Experience
50	Communicate psychometric analysis results	Documentation
51	Knowledge of classical test theory and item response theory	Psychometrics
52	Develop and document psychometric solutions	Documentation; Psychometrics
53	Communicate technical concepts in non-technical ways	Documentation
54	Synthesize information & produce coherent summaries & recommendations	Documentation; Consultation
55	Create documents, prepare presentations, and present to stakeholders	Documentation
56	Articulate legislative testing requirements	Experience
57	Lead standard setting activities	Consultation
58	Train item-writers to write items	Consultation
59	Facilitate review meetings for item analyses and other analyses	Consultation; Psychometrics; Documentation

Appendix B. Psychometrician Exam Blueprint

Topic ID	**Topic**	**Weight**	**Description**
#1	***Data Management and Analysis***	***28***	
#1.1	Planning analyses	8	Knowledge of: analytic methods and feasibility; sample size requirements and statistical power; data collection design and methods. Ability to document research questions and hypotheses.
#1.2	Conducting analyses	8	Knowledge of statistical concepts and hypothesis testing. Application of statistical methods to analyze data and interpret correctly to reach generalizable empirical conclusions
#1.3	Wrangle data	4	Assemble, order, recode, join, subset, and store datasets for analysis.
#1.4	Ensure data quality	4	Maintain data quality by examining data, using exploratory data analysis, removing outliers, cleaning data, etc.
#1.5	Scripting in R	2	Scripting in R to accomplish a step in research, analysis, reporting, etc.

Topic ID	Topic	Weight	Description
#1.6	SQL Queries	2	Writing SQL to query data; focus on features a psychometrician would use and issues they might encounter like SELECT, joins, grouping, counting, missing data, etc.
#2	***Consultation***	***10***	
#2.1	Advising about exam development	4	Consult with stakeholders about exam development, including: high-level decisions about scope, methodology, etc.; steps of exam development; advice about fitting program needs to exam development methods
#2.2	SME training and facilitation	6	Support subject matter experts in all stages of exam development with training, facilitation, coaching, etc.
#3	***Reporting and Documentation***	***8***	
#3.1	Write reports and papers	4	Write reports of statistical analyses, including: reviewing relevant research literature, framing research questions, describing methodology, communicating results, interpreting results, integrating results with prior research, providing conclusions. Authors reports tailored to the audience. Writes and publishes research articles in industry journals.
#3.2	Present results	4	Prepares presentations summarizing the key findings and conclusions. Presents results to stakeholders, peers, and the general population.
#4	***Knowledge of Psychometric Practice***	***10***	
#4.1	Industry knowledge	2	Broad knowledge of the testing industry, including: exam programs, potential clients and partners, competitors, notable professionals, regulators and regulations.
#4.2	Knowledge of ethical and legal standards	4	Knowledge of best practices in developing exams. Includes specialized knowledge about best practices within specific domains (clinical, educational, organizational, credentialing, etc.). Knowledge of successful solutions and models that have worked in the past that can be adapted for new circumstances.
#4.3	Knowledge of psychometric standards	4	Knowledge of psychometric standards
#5	***Knowledge of Psychometric Theory***	***18***	
#5.1	Knowledge of test theories	6	Knowledge and application of: reliability and validity; classical test theory; latent trait and IRT;
#5.2	Knowledge of scaling and equating	6	Knowledge and application of: different scale types and their associated level of measurement; standardized scores; CTT and IRT methods for equating test scores
#5.3	Knowledge of exam analyses	6	Knowledge and application of: item analysis including difficulty and quality; scale/exam analysis including score distribution, dimensionality, reliability, passing scores

Topic ID	Topic	Weight	Description
#6	***Research***	***20***	
#6.1	Conducts job analyses	4	Knowledge of the purpose and steps involved in a job analysis: collect information (tasks, skills, or competencies) about the position; evaluate the criticality of each task/skill/competency; research industry standards; use collected data to create job description/exam blueprint based on the JTA
#6.2	Conducts standard setting	4	Select the appropriate standard setting method; design materials needed in standard setting; identify SMEs qualified for the study; use collected data to calculate the cut-score and estimate the error associated with it
#6.3	Conducts validation study	4	Knowledge of different types of validity (content, criterion-related, etc.) study and their purposes;
#6.4	Conducts fairness research	4	Conducts fairness research
#6.5	Conducts other applied research	4	Conducts other applied research
#7	***Software***	***6***	
#7.1	Installing and using analysis software	4	Installing and using analysis software
#7.2	Designing psychometric software	2	Designing psychometric software

Appendix C. Specific Topics Used for Item Generation

Item	Topic ID	Specific Topic
1	#1.1	sample size planning
2	#1.1	sample size planning
3	#1.1	statistical power and type I error
4	#1.1	statistical power and type I error
5	#1.1	analytic methods for nominal data
6	#1.1	analytic methods for nominal data
7	#1.1	circumstances where non-parametric hypothesis tests should be used
8	#1.1	circumstances where non-parametric hypothesis tests should be used
9	#1.1	non-parametric hypothesis tests and statistical power
10	#1.1	non-parametric hypothesis tests and statistical power
11	#1.2	how is Cohen's d = 0.40 interpreted?
12	#1.2	how is Cohen's d = 0.40 interpreted?
13	#1.2	what are the assumptions of an independent sample t-test

Item	Topic ID	Specific Topic
14	#1.2	what are the assumptions of an independent sample t-test
15	#1.2	when should logistic regression be used instead of linear regression
16	#1.2	choice of logistic or linear regression analysis
17	#2.1	how many cases are needed for a job task analysis survey
18	#2.1	what are sample size requirements for a job-task analysis survey
19	#2.1	how often should a JTA be repeated
20	#2.1	when should you repeat a JTA for a given job
21	#2.1	how many items do you need on a typical IT certification exam
22	#2.1	exam length, in items, for a typical certification exam
23	#2.2	the most important element of writing effective multiple-choice exam items
24	#2.2	the most important element of writing effective multiple-choice exam items
25	#2.2	how should subject matter experts write JTA task statements
26	#2.2	ideal JTA task statements for knowledge domains
27	#2.2	rating scales for JTA surveys
28	#2.2	rating scales for JTA surveys
29	#3.1	scoring Likert items
30	#3.1	scoring Likert items for reliability analyses
31	#3.1	treating missing data for omitted items on multiple choice knowledge-based exams
32	#3.1	how should missing data for omitted items on multiple choice knowledge-based exams be scored when scoring items
33	#3.1	how should items on a knowledge exam be scored if the candidate ran out of time
34	#3.1	scoring not-reached item responses on a knowledge-based multiple-choice exam
35	#3.1	steps in cleaning exam data
36	#3.1	detecting bivariate outliers
37	#3.2	criteria for rejecting cases in examination data
38	#3.2	criteria for rejecting cases in examination data
39	#3.2	detecting cheating on exams with correct and incorrect responses
40	#3.2	detecting cheating on exams with correct and incorrect responses
41	#6.1	classical test theory correction for attenuation
42	#6.1	strengths and weaknesses of methods for estimating psychometric reliability
43	#6.1	IRT conditional standard error
44	#6.1	IRT assumption of unidimensionality
45	#6.1	testing the IRT assumption of unidimensionality

Appendix D. Rater Estimated Usability of the Items

Item ID	Item Usability (text-davinci-002)				Item Usability (text-davinci-003)		
	Average	Rater 1	Rater 2		Average	Rater 1	Rater 2
1	2	2	2		2	2	2
2	2.5	2	3		2.5	2	3
3	2.5	3	2		2	2	2
4	2	2	2		2.5	3	2
5	1.5	1	2		3.5	4	3
6	1	1	1		3	3	3
7	3	3	3		2.5	2	3
8	2	2	2		2	2	2
9	1	1	1		2	2	2
10	2	2	2		1	1	1
11	3	3	3		3	3	3
12	3	3	3		3.5	4	3
13	3.5	3	4		3	3	3
14	4	4	4		2.5	2	3
15	3	3	3		3	3	3
16	3.5	3	4		1	1	1
17	2	2	2		2	2	2
18	3	3	3		2	2	2
19	2.5	4	1		2.5	2	3
20	3	3	3		2.5	3	2
21	2.5	3	2		2.5	3	2
22	3	2	4		2.5	3	2
23	2	2	2		3	3	3
24	2	2	2		2.5	2	3
25	1.5	1	2		2.5	3	2
26	2.5	3	2		1.5	1	2
27	1	1	1		1.5	1	2

Item ID	Item Usability (text-davinci-002)				Item Usability (text-davinci-003)		
	Average	Rater 1	Rater 2		Average	Rater 1	Rater 2
28	2	1	3		1	1	1
29	1	1	1		1.5	1	2
30	2	2	2		1.5	1	2
31	2	2	2		2.5	3	2
32	3	3	3		3	3	3
33	3	3	3		3	3	3
34	1.5	2	1		3	3	3
35	1.5	2	1		2	2	2
36	3.5	4	3		3	3	3
37	1.5	1	2		2	2	2
38	2.5	2	3		2.5	2	3
39	1	1	1		2.5	3	2
40	1.5	2	1		2.5	3	2
41	2	3	1		1	1	1
42	1.5	2	1		2	2	2
43	2	2	2		1.5	2	1
44	2.5	2	3		3	3	3
45	1.5	2	1		3.5	4	3

Note. Likert-scale used for Item Usability ratings: 1 = Useless; 2 = Needs much work; 3 = Needs minor edits; 4 = Acceptable as is.

Appendix E. Items generated

This appendix presents the full text of the items generated. For example, the first item for the "text-davinci-002" model was:
Path: ./psy-002
Item 1:

power, alpha, beta, and effect size.

You are planning to conduct a study to compare the effectiveness of two different teaching methods. Which of the following would be the best sample size to use in order to have enough power to detect a small difference between the two methods?
A. N = 2

B. N = 10
C. N = 50
D. N = 100
Correct Answer: D. N = 100
Topic ID:1.1
(finish reason = stop; usage = 416 tokens)
The bolded text is added for reference. The italicized text is the verbatim response (including, for this item, the stray list of words at the beginning and the three newlines after this list).

psych-items-002

Path: ./psy-002
Item 1:
power, alpha, beta, and effect size.

You are planning to conduct a study to compare the effectiveness of two different teaching methods. Which of the following would be the best sample size to use in order to have enough power to detect a small difference between the two methods?
A. N = 2
B. N = 10
C. N = 50
D. N = 100

Correct Answer: D. N = 100
Topic ID:1.1
(finish reason = stop; usage = 416 tokens)

Path: ./psy-002
Item 2:

You are a psychometrician working on a new assessment. You need to determine the sample size for the assessment. Which of the following would be the most important factor to consider when determining sample size?

A. The number of items on the assessment
B. The desired level of precision for the results
C. The number of test-takers
D. The amount of time available to administer the assessment

Correct Answer: B
Topic ID:1.1
(finish reason = stop; usage = 418 tokens)

Path: ./psy-002
Item 3:

You are working on a research project and want to ensure that you have enough statistical power to detect an effect. Which of the following would be the MOST important factor to consider?
A. The sample size
B. The type I error
C. The type II error
D. The effect size

Correct Answer: B
Topic ID:1.1
(finish reason = stop; usage = 400 tokens)

Path: ./psy-002
Item 4:

You are planning to run an analysis on a new data set. What are the three main things you need to consider when determining the most appropriate statistical test to use?
A. The type of data collected
B. The number of variables in the data set
C. The statistical power and type I error
Correct Answer: C
Topic ID:1.1
(finish reason = stop; usage = 400 tokens)

Path: ./psy-002
Item 5:
frequencies, percentages, means, standard deviations, correlations, and Chi-square analyses.

You are working on a research project that requires you to analyze data collected from a survey. Which of the following methods would be the most appropriate for analyzing the data?

A. Linear regression
B. Logistic regression
C. Correlation
D. Chi-square analyses

Correct Answer: D. Chi-square analyses
Topic ID:1.1
(finish reason = stop; usage = 414 tokens)

Path: ./psy-002
Item 6:

You are planning to analyze the data collected from a study. Which of the following methods would be the most appropriate for this type of data?
A. Mean
B. Median
C. Mode
D. Standard deviation

Correct Answer: D. Standard deviation
Topic ID:1.1
(finish reason = stop; usage = 384 tokens)

Path: ./psy-002
Item 7:

You are planning to conduct a study to compare the mean number of hours that full-time employees work in different countries. Which of the following statements is true?
A. Non-parametric hypothesis tests should be used when the data are not normally distributed.
B. Non-parametric hypothesis tests should be used when the population variance is unknown.
C. Non-parametric hypothesis tests should be used when the sample size is small.
D. Non-parametric hypothesis tests should be used when the data are ordinal.

Correct Answer: B
Topic ID:1.1
(finish reason = stop; usage = 450 tokens)

Path: ./psy-002
Item 8:
You are administering a test to a sample of 100 individuals. You want to use a parametric hypothesis test to test whether the population mean is different from 10. Which of the following do you need to be true in order for the parametric hypothesis test to be appropriate?
A. The population is normally distributed.
B. The population variance is known.
C. The sample size is small.
D. The sample size is large.

Correct Answer: A
Topic ID:1.1
(finish reason = stop; usage = 430 tokens)

Path: ./psy-002
Item 9:

You are working on a research project and want to know which non-parametric hypothesis test to use. You also want to know how large your sample size should be in order to have sufficient statistical power. Which of the following would be the best resource to consult?

A. Your textbook on introductory statistics
B. A statistics journal article on non-parametric hypothesis tests
C. The statistical software manual
D. The research proposal guidelines

Correct Answer: B
Topic ID:1.1
(finish reason = stop; usage = 429 tokens)

Path: ./psy-002
Item 10:

You are planning to conduct a study to compare the mean score of two groups on a new intelligence test. You want to be sure that your study has enough statistical power to detect a small but meaningful difference between the groups. What sample size should you use?

A. 50 participants in each group
B. 100 participants in each group
C. 150 participants in each group
D. 200 participants in each group

Correct Answer: B
Topic ID:1.1
(finish reason = stop; usage = 424 tokens)

Path: ./psy-002
Item 11:

You are working as a psychometrician. Your boss asks you to explain how Cohen's d = 0.40 is interpreted. You say:

A. Cohen's d = 0.40 is interpreted as the average effect size in the population.
B. Cohen's d = 0.40 is interpreted as the difference between the means of two groups.
C. Cohen's d = 0.40 is interpreted as the percentage of variance explained by the difference between two groups.
D. Cohen's d = 0.40 is interpreted as the correlation between two variables.

Correct Answer: B. Cohen's d = 0.40 is interpreted as the difference between the means of two groups.
Topic ID:1.2
(finish reason = stop; usage = 459 tokens)

Path: ./psy-002
Item 12:

You are conducting analyses for a new study. Which of the following is true about Cohen's d = 0.40?

A. It means that the difference between the two groups is not statistically significant.
B. It means that the difference between the two groups is statistically significant.
C. It means that the difference between the two groups is small.
D. It means that the difference between the two groups is large.

Correct Answer: C
Topic ID:1.2
(finish reason = stop; usage = 412 tokens)

Path: ./psy-002
Item 13:

You are conducting a study to compare the mean scores of two groups on a new assessment. You want to use an independent sample t-test. Which of the following is not an assumption of the independent sample t-test?
A. The groups are independent
B. The dependent variable is interval
C. The dependent variable is normally distributed
D. The variances of the two groups are equal

Correct Answer: D. The variances of the two groups are equal.
Topic ID:1.2
(finish reason = stop; usage = 416 tokens)

Path: ./psy-002
Item 14:

A researcher is interested in knowing whether there is a difference in the mean scores of two different groups on a particular measure. She decides to use an independent sample t-test to analyze her data. Which of the following is NOT an assumption of the independent sample t-test?
A) The two groups have equal variances
B) The two groups have equal means
C) The two groups are independent of each other
D) The dependent variable is normally distributed

Correct Answer: B
Topic ID:1.2
(finish reason = stop; usage = 418 tokens)

Path: ./psy-002
Item 15:

Scenario:

You are working as a psychometrician on a project to assess the technical quality of educational assessments. Part of your job is to lead operational and research projects, ensure quality control of deliverables, and contribute to the development of data interpretation materials and publications. In order to do this effectively, it is important to have a strong understanding of when to use logistic regression instead of linear regression.

Which of the following is a correct statement about when to use logistic regression instead of linear regression?

A. Logistic regression should be used when the outcome variable is dichotomous.
B. Logistic regression should be used when the outcome variable is continuous.
C. Logistic regression should be used when the predictor variables are dichotomous.
D. Logistic regression should be used when the predictor variables are continuous.

Correct Answer: A
Topic ID:1.2
(finish reason = stop; usage = 497 tokens)

Path: ./psy-002
Item 16:

Scenario:

You are working on a research project that aims to predict whether or not a student will pass their final exams. You have collected data on a number of variables, including the student's grade point average, the number of hours spent studying, and the difficulty of the course. You want to use this data to develop a model that can be used to predict whether or not a student will pass their final exams.

Which of the following methods would be the most appropriate for this task?

A. Linear regression
B. Logistic regression

C. Multiple regression
D. Chi-squared test

Correct Answer: B. Logistic regression
Topic ID:1.2
(finish reason = stop; usage = 451 tokens)

Path: ./psy-002
Item 17:

You are a psychometrician consultant who has been asked to advise a client on how many cases are needed for a job task analysis survey. Which of the following is the best advice to give your client?

A. It is not necessary to have any cases for a job task analysis survey.
B. A minimum of 10 cases is necessary for a job task analysis survey.
C. A minimum of 30 cases is necessary for a job task analysis survey.
D. A minimum of 50 cases is necessary for a job task analysis survey.

Correct Answer: C
Topic ID:2.1
(finish reason = stop; usage = 459 tokens)

Path: ./psy-002
Item 18:

You are a psychometrician working on a new exam. You consult with stakeholders about the development of the exam, including the scope, methodology, and steps of development. One of the stakeholders asks you about the sample size requirements for a job-task analysis survey. You respond:

A. The sample size requirements for a job-task analysis survey depend on the number of job tasks being analyzed.
B. The sample size requirements for a job-task analysis survey depend on the number of items on the exam.
C. A job-task analysis survey should include all tasks being performed by the job incumbents.
D. The sample size requirements for a job-task analysis survey depend on the number of job incumbents being surveyed.

Correct Answer: D
Topic ID:2.1
(finish reason = stop; usage = 501 tokens)

Path: ./psy-002
Item 19:

A. Once every five years.
B. As needed, but at least once every five years.
C. As needed, but at least once every three years.
D. As needed, but at least once every two years.

Correct Answer: B
Topic ID:2.1
(finish reason = stop; usage = 394 tokens)

Path: ./psy-002
Item 20:

You are a psychometrician working on the development of an exam for a new job position. You have consulted with the stakeholders and have determined that a job task analysis (JTA) is necessary. You are now ready to begin the process of writing items for the exam.

Which of the following is true about when you should repeat a JTA for a given job?

A. You should repeat a JTA every time a new job position is created.
B. You should repeat a JTA whenever the job duties for a given position change.
C. You should repeat a JTA whenever the performance standards for a given position change.
D. You should not repeat a JTA unless the content of the job position has changed significantly.

Correct Answer: B
Topic ID:2.1
(finish reason = stop; usage = 502 tokens)

Path: ./psy-002
Item 21:

You are a psychometrician working on an IT certification exam. You consult with the stakeholders about the development of the exam. They would like to know how many items you need on the exam. What do you tell them?

A. As few as possible, to reduce time and costs
B. Enough to cover the content domain
C. A minimum of 100 items
D. The number of items needed is based on the size of the content domain and the level of difficulty required for the test-takers

Correct Answer: D
Topic ID:2.1
(finish reason = stop; usage = 454 tokens)

Path: ./psy-002
Item 22:

You are a psychometrician consulting with a client about their upcoming certification exam. They are interested in knowing how long the exam should be, in terms of the number of items. What would you advise them?

A. The exam should be as long as possible, to ensure that all content areas are covered.
B. The exam should be as short as possible, to minimize the amount of time test-takers have to spend on the exam.
C. The exam should be of moderate length, to strike a balance between coverage of content areas and time spent on the exam.
D. The length of the exam should be based on the number of content areas that need to be covered.

Correct Answer: C. The exam should be of moderate length, to strike a balance between coverage of content areas and time spent on the exam.
Topic ID:2.1
(finish reason = stop; usage = 516 tokens)

Path: ./psy-002
Item 23:

Which of the following is the most important element of writing effective multiple-choice exam items?

A. Wording of the stem
B. Wording of the distractors
C. Placement of the correct answer
D. All of the above

Correct Answer: B. Wording of the distractors
Topic ID:2.2
(finish reason = stop; usage = 396 tokens)

Path: ./psy-002
Item 24:

What is the most important element of writing effective multiple-choice exam items?

A. making sure that all answer choices are approximately equal in length
B. making sure that the answer choices are mutually exclusive
C. making sure that the answer choices are exhaustive
D. making sure that the answer choices are clearly and concisely worded

Correct Answer: B
Topic ID:2.2
(finish reason = stop; usage = 407 tokens)

Path: ./psy-002
Item 25:

What is the first step that a subject matter expert (SME) should take when writing a job task analysis (JTA) statement?

A. Review the JTA standards.
B. Select the JTA tasks.
C. Write the JTA task statements.
D. Pretest the JTA tasks.

Correct Answer: B. Select the JTA tasks.
Topic ID:2.2
(finish reason = stop; usage = 411 tokens)

Path: ./psy-002
Item 26:

You are working with a team of SMEs on developing a new assessment. Which of the following would be the MOST important task statement for a knowledge domain?

A. To understand and be able to apply the concepts of the knowledge domain
B. To be able to explain the theories and models of the knowledge domain

C. To be able to use the tools and methods of the knowledge domain
D. To be able to apply the knowledge domain to new situations

Correct Answer: D
Topic ID:2.2
(finish reason = stop; usage = 428 tokens)

Path: ./psy-002
Item 27:

You are a psychometrician working on a job task analysis (JTA) survey. You need to choose a rating scale for the JTA surveys. Which of the following rating scales would you use?

A. 1-5 scale, with 1 indicating 'Strongly Disagree' and 5 indicating 'Strongly Agree'
B. 1-7 scale, with 1 indicating 'Strongly Disagree' and 7 indicating 'Strongly Agree'
C. 1-9 scale, with 1 indicating 'Strongly Disagree' and 9 indicating 'Strongly Agree'
D. 1-10 scale, with 1 indicating 'Strongly Disagree' and 10 indicating 'Strongly Agree'

Correct Answer: B
Topic ID:2.2
(finish reason = stop; usage = 477 tokens)

Path: ./psy-002
Item 28:

You are a psychometrician working on a new job analysis survey. You need to choose a rating scale for the survey. Which of the following would be the most appropriate rating scale to use?

A. Likert Scale
B. Guttman Scale
C. Thurstone Scale
D. Semantic Differential Scale

Correct Answer: A
Topic ID:2.2
(finish reason = stop; usage = 400 tokens)

Path: ./psy-002
Item 29:

You are a psychometrician working on an exam. You need to score the following likert items:

1. I am very interested in data management
2. I am skilled in data management
3. I am familiar with data management
4. I have no interest in data management

Correct Answer: B
Topic ID:3.1
(finish reason = stop; usage = 384 tokens)

Path: ./psy-002
Item 30:

Scoring likert items for reliability analyses:

A. Is a process of assigning numerical values to responses on a scale in order to measure variables B. Is a process of estimating the internal consistency of a scale C. Is a method of deriving a measure of central tendency for a set of ordinal data D. Is a process of creating a new set of data by combining two or more existing sets of data

Correct Answer: B
Topic ID:3.1
(finish reason = stop; usage = 411 tokens)

Path: ./psy-002
Item 31:

Which of the following is the best way to treat missing data for omitted items on multiple choice knowledge-based exams?

A. Estimate the missing values using the average of the responses to the other items on the test.
B. Estimate the missing values using the average of the responses to the other items in the same content area on the test.
C. Estimate the missing values using the average of the responses to the other items in the same content area on other tests.
D. Do not estimate the missing values.

Correct Answer: D
Topic ID:3.1
(finish reason = stop; usage = 443 tokens)

Path: ./psy-002
Item 32:

Which of the following is the best way to score missing data for omitted items on multiple choice knowledge-based exams?

A. Omit the item from the analysis
B. Score the item as 0
C. Score the item as 1
D. Score the item according to the answer key

Correct Answer: D
Topic ID:3.1
(finish reason = stop; usage = 400 tokens)

Path: ./psy-002
Item 33:

You are a psychometrician working on a knowledge exam. If a candidate runs out of time, how should the items be scored?

A. The items should be left blank and not scored.
B. The items should be scored as if the candidate had answered them correctly.
C. The items should be scored as if the candidate had answered them incorrectly.
D. The items should be randomly scored.

Correct Answer: A. The items should be left blank and not scored.
Topic ID:3.1
(finish reason = stop; usage = 436 tokens)

Path: ./psy-002
Item 34:

You are developing a knowledge-based multiple-choice exam. One item on the exam asks the test takers to identify the most important reason for scoring not-reached item responses. Which of the following is the best reason for scoring not-reached item responses?
A. To ensure the fairness of the exam
B. To ensure the validity of the exam
C. To ensure the reliability of the exam
D. To ensure the objectivity of the exam

Correct Answer: B
Topic ID:3.1
(finish reason = stop; usage = 431 tokens)

Path: ./psy-002
Item 35:
recode, subset, and join dataframes.

You are given an exam data set that you need to clean for analysis. Which of the following is the correct order of steps you should take?

A. Join dataframes, subset dataframes, recode dataframes.
B. Recode dataframes, join dataframes, subset dataframes.
C. Subset dataframes, join dataframes, recode dataframes.
D. Join dataframes, recode dataframes, subset dataframes.

Correct Answer: D
Topic ID:3.1
(finish reason = stop; usage = 430 tokens)

Path: ./psy-002
Item 36:

Which of the following best describes how to detect bivariate outliers?
A. By looking at the scatterplot of the data
B. By looking at the histogram of the data
C. By looking at the boxplot of the data
D. By looking at the summary statistics of the data

Correct Answer: A
Topic ID:3.1
(finish reason = stop; usage = 387 tokens)

Path: ./psy-002
Item 37:

You are reviewing the examination data for your upcoming psychometrics exam and notice that some of the cases are outliers. You decide to reject these cases from the data. What is the best reason to reject these cases?

A. To ensure the data quality
B. To increase the validity of the data
C. To reduce the variability of the data
D. To improve the reliability of the data

Correct Answer: A
Topic ID:3.2
(finish reason = stop; usage = 416 tokens)

Path: ./psy-002
Item 38:

A psychometrician is reviewing examination data and wants to ensure data quality. Which of the following is NOT a criterion for rejecting cases in examination data?
A. The cases are not valid responses to the items.
B. The cases have missing values.
C. The cases are outliers.
D. The cases have been randomly selected.

Correct Answer: D
Topic ID:3.2
(finish reason = stop; usage = 405 tokens)

Path: ./psy-002
Item 39:

You are a psychometrician working on an exam. You notice that there are a lot of correct and incorrect responses. What do you do?

A. Ignore the data and move on.
B. Report the data to your supervisor.
C. Examine the data to see if there is any cheating going on.
D. Clean the data to remove any outliers.

Correct Answer: C. Examine the data to see if there is any cheating going on.
Topic ID:3.2
(finish reason = stop; usage = 429 tokens)

Path: ./psy-002
Item 40:

You are a psychometrician working on an exam. You notice that there are an unusually high number of correct and incorrect responses. What is the most likely explanation?

A) The test is too easy and needs to be made more difficult.
B) The test is too difficult and needs to be made easier.

C) Cheating is occurring on the exam.
D) The test is adequate and no changes need to be made.

Correct Answer: C) Cheating is occurring on the exam.
Topic ID:3.2
(finish reason = stop; usage = 435 tokens)

Path: ./psy-002
Item 41:

A psychometrician is conducting a study to determine the reliability of a new test. Which of the following would be the best estimate of the reliability of the test?
A. The percentage of variance in the test scores that is due to true differences in the abilities of the test takers.
B. The percentage of variance in the test scores that is due to error.
C. The correlation between the test scores and the scores on an established, similar test.
D. The percentage of test takers who achieve the passing score on the test.

Correct Answer: B
Topic ID:6.1
(finish reason = stop; usage = 431 tokens)

Path: ./psy-002
Item 42:
alternate forms, split-half, and internal consistency.

Which of the following methods for estimating psychometric reliability is the best?
A. Alternate forms
B. Split-half
C. Internal consistency

Correct Answer: A. Alternate forms
Topic ID:6.1
(finish reason = stop; usage = 365 tokens)

Path: ./psy-002
Item 43:

You are asked to review the psychometric properties of a new adaptive testing program. Which of the following is NOT a true statement about the IRT conditional standard error?
A. It is the error that would be expected if a person were to take the same test an infinite number of times.
B. It is the error that would be expected if a person were to take the same test twice.
C. It is the error that would be expected if a person were to take the same test at two different points in time.
D. It is the error that would be expected if a person were to take two different versions of the same test.

Correct Answer: D
Topic ID:6.1
(finish reason = stop; usage = 449 tokens)

Path: ./psy-002
Item 44:
 the idea that a single latent construct underlies the responses to all items on a test.

Which of the following is an IRT assumption of unidimensionality?
A. The idea that a single latent construct underlies the responses to all items on a test.
B. The idea that a test measures more than one latent construct.
C. The idea that a test measures only a single latent construct.
D. The idea that more than one latent construct underlies the responses to all items on a test.

Correct Answer: A
Topic ID:6.1
(finish reason = stop; usage = 427 tokens)

Path: ./psy-002
Item 45:
 what it is, how to assess it, and how to remedy it when it's not met.

You are working on a new test. You want to make sure the items on the test are measuring the same thing. What is the name of the statistical assumption you are testing for?
A. Unidimensionality
B. Reliability
C. Validity
D. Normality

Correct Answer: A. Unidimensionality
Topic ID:6.1
(finish reason = stop; usage = 411 tokens)

psych-items-003

Path: ./psy-003
Item 1:

Which of the following is a key factor to consider when determining a sample size for a psychometric study?
A. The cost of the research
B. The amount of statistical power needed
C. The complexity of the research question
D. The amount of data collected

Correct Answer: B. The amount of statistical power needed
Topic ID:1.1
(finish reason = stop; usage = 398 tokens)

Path: ./psy-003
Item 2:

A researcher is planning a study and needs to determine the sample size. Which of the following is the most important factor in determining sample size?

A. The number of items in the survey
B. The anticipated response rate
C. The level of statistical power desired
D. The number of groups in the study

Correct Answer: C. The level of statistical power desired
Topic ID:1.1
(finish reason = stop; usage = 405 tokens)

Path: ./psy-003
Item 3:

When setting up a study, a Psychometrician needs to understand statistical power and type I error in order to:
A. Accurately interpret results.
B. Set up the sample size requirements.
C. Estimate the analysis methods and feasibility.
D. Estimate data collection design and methods.

Correct Answer: B. Set up the sample size requirements.
Topic ID:1.1
(finish reason = stop; usage = 410 tokens)

Path: ./psy-003
Item 4:

Scenario: An individual is planning to conduct an experiment to compare the effects of two treatments on a research population. Which of the following is the most important factor for determining the statistical power and the type I error for the experiment?
A. The number of participants in the study
B. The size of the effect being studied
C. The size of the alpha level
D. The reliability of the measurement instrument

Correct Answer: A
Topic ID:1.1
(finish reason = stop; usage = 422 tokens)

Path: ./psy-003
Item 5:

Which of the following analytical methods is best suited for nominal data?
A. Linear regression
B. Logistic regression
C. Chi-square
D. Factor analysis

Correct Answer: C. Chi-square
Topic ID:1.1
(finish reason = stop; usage = 374 tokens)

Path: ./psy-003
Item 6:

Scenario: You are a Psychometrician working on a project and need to analyze nominal data. What would be the most appropriate analytic method to use?
A. Correlation
B. Chi-Square Test
C. Regression
D. ANOVA

Correct Answer: B. Chi-Square Test
Topic ID:1.1
(finish reason = stop; usage = 393 tokens)

Path: ./psy-003
Item 7:

A Psychometrician is planning analyses of test results. Under which of the following circumstances should non-parametric hypothesis tests be used?

A. When the data is distributed in a bell-shaped curve
B. When the data is categorical
C. When the data is ordinal
D. When the data is normally distributed

Correct Answer: B. When the data is categorical
Topic ID:1.1
(finish reason = stop; usage = 417 tokens)

Path: ./psy-003
Item 8:

When determining the best statistical test for a particular research question, which of the following circumstances would indicate that a non-parametric hypothesis test is the most appropriate?

A. When the data is normally distributed
B. When the population variance is known
C. When the data is ordinal or categorical
D. When the sample size is small

Correct Answer: C. When the data is ordinal or categorical
Topic ID:1.1
(finish reason = stop; usage = 424 tokens)

Path: ./psy-003
Item 9:

When planning a study to compare the results of two different treatments, which of the following hypothesis tests is most appropriate for the analysis and why?
A. Mann-Whitney U-test, because it is a non-parametric test.
B. Chi-square, because it is a non-parametric test.
C. t-test, because it is a parametric test.

D. ANOVA, because it has higher statistical power.

Correct Answer: A. Mann-Whitney U-test, because it is a non-parametric test.
Topic ID:1.1
(finish reason = stop; usage = 452 tokens)

Path: ./psy-003
Item 10:

Scenario: You are a Psychometrician and need to determine which approach to use for a non-parametric hypothesis test.
A. Linear regression
B. Linear correlation
C. Chi-square
D. Mann-Whitney U test

Correct Answer: D. Mann-Whitney U test
Topic ID:1.1
(finish reason = stop; usage = 397 tokens)

Path: ./psy-003
Item 11:

Scenario: A Psychometrician is conducting a study to determine the effect of an intervention on a certain population. After analyzing the data, they determine that the mean difference between the two groups is 0.40 standard deviations. How is this mean difference interpreted?

A. The intervention had no significant effect on the population
B. The intervention had a statistically significant effect on the population
C. The intervention had a small effect on the population
D. The intervention had a large effect on the population

Correct Answer: B. The intervention had a statistically significant effect on the population
Topic ID:1.2
(finish reason = stop; usage = 435 tokens)

Path: ./psy-003
Item 12:
Which of the following is the most accurate interpretation of Cohen's d = 0.40?
A. There is a small to moderate effect size.
B. There is a large effect size.
C. There is no effect size.
D. There is a moderate to large effect size.

Correct Answer: A. There is a small to moderate effect size.
Topic ID:1.2
(finish reason = stop; usage = 394 tokens)

Path: ./psy-003
Item 13:

Scenario: You are working on a research project and are considering using an independent sample t-test as your analysis technique. What are the assumptions for this test?

A. The samples must have equal variance
B. The population variance must be known
C. The data must be normally distributed
D. The samples must be independent

Correct Answer: D. The samples must be independent
Topic ID:1.2
(finish reason = stop; usage = 398 tokens)

Path: ./psy-003
Item 14:
Which of the following are assumptions of an Independent Sample t-test?
A. Samples are randomly selected from the same population
B. Variances of the two samples are equal
C. Samples are normally distributed
D. Samples are independent of each other

Correct Answer: B, D
Topic ID:1.2
(finish reason = stop; usage = 385 tokens)

Path: ./psy-003
Item 15:

Scenario: A Psychometrician is analyzing data collected from a test. When should logistic regression be used instead of linear regression?

A) When the data is categorical in nature
B) When the data is continuous in nature
C) When the data is dichotomous in nature
D) When the data is ordinal in nature

Correct Answer: A) When the data is categorical in nature
Topic ID:1.2
(finish reason = stop; usage = 404 tokens)

Path: ./psy-003
Item 16:

Scenario: You are conducting a research project which requires analyzing the data collected. What type of analysis should you use?
a. Standardized Test Theory
b. Factor Analysis
c. Logistic Regression
d. Linear Regression

Correct Answer: d. Linear Regression
Topic ID:1.2

(finish reason = stop; usage = 373 tokens)

Path: ./psy-003
Item 17:

Which of the following is the best estimate of the number of cases needed for a job task analysis survey?
A. 10-20
B. 20-30
C. 30-40
D. 40-50

Correct Answer: B. 20-30
Topic ID:2.1
(finish reason = stop; usage = 397 tokens)

Path: ./psy-003
Item 18:

When developing an exam, a Psychometrician must consult with stakeholders about the sample size requirements for a job-task analysis survey. Which of the following is the correct size of the sample for this survey?

A. 100
B. 200
C. 500
D. 1,000

Correct Answer: D. 1,000
Topic ID:2.1
(finish reason = stop; usage = 416 tokens)

Path: ./psy-003
Item 19:

Which of the following best describes how often a Job Task Analysis (JTA) should be repeated?
A. Whenever the job tasks change
B. Every year
C. Every 3-5 years
D. Once, as long as the job tasks remain the same

Correct Answer: C. Every 3-5 years
Topic ID:2.1
(finish reason = stop; usage = 406 tokens)

Path: ./psy-003
Item 20:

Which of the following is the best point in time to repeat a job task analysis (JTA) for a given job?

a. After the job has changed significantly
b. Prior to item writing
c. After each exam cycle
d. After the job has been in existence for two years

Correct Answer: a. After the job has changed significantly
Topic ID:2.1
(finish reason = stop; usage = 416 tokens)

Path: ./psy-003
Item 21:

Which of the following best reflects the number of items needed on a typical IT certification exam?
A. Between 10 and 20
B. Between 20 and 30
C. Between 30 and 40
D. Over 40

Correct Answer: D. Over 40
Topic ID:2.1
(finish reason = stop; usage = 396 tokens)

Path: ./psy-003
Item 22:

Which of the following is the most typical length, in items, for a certification exam?
A. 25 items
B. 50 items
C. 100 items
D. 150 items

Correct Answer: C. 100 items
Topic ID:2.1
(finish reason = stop; usage = 389 tokens)

Path: ./psy-003
Item 23:

Which of the following is the most important element of writing effective multiple-choice exam items?
A. Crafting clear and unambiguous stems
B. Ensuring that distractors are plausible
C. Writing challenging items
D. Training and facilitation of SMEs

Correct Answer: A. Crafting clear and unambiguous stems
Topic ID:2.2
(finish reason = stop; usage = 398 tokens)

Path: ./psy-003
Item 24:

Scenario: You are a Psychometrician tasked with SME training and facilitation. What is the most important element of writing effective multiple-choice exam items?

A. Ensuring that the items are valid and reliable
B. Writing items that are clear, unambiguous, and easy to understand
C. Writing items that are complex and difficult to answer
D. Writing items that are relevant to the subject matter

Correct Answer: B. Writing items that are clear, unambiguous, and easy to understand
Topic ID:2.2
(finish reason = stop; usage = 438 tokens)

Path: ./psy-003
Item 25:

Which of the following is the most effective way for subject matter experts to write job task analysis (JTA) task statements?
a. Use jargon and technical terms
b. Use simple language and keep sentences direct and concise
c. Use long, complex sentences
d. Use complex language and include details

Correct Answer: b. Use simple language and keep sentences direct and concise
Topic ID:2.2
(finish reason = stop; usage = 409 tokens)

Path: ./psy-003
Item 26:

Which of the following is an ideal JTA task statement for a knowledge domain?
A. Creating and maintaining an exam blueprint
B. Ensuring quality control of deliverables
C. Identifying the educational and experiential requirements of a job
D. Setting performance standards and cutscores

Correct Answer: C. Identifying the educational and experiential requirements of a job
Topic ID:2.2
(finish reason = stop; usage = 406 tokens)

Path: ./psy-003
Item 27:

Which of the following is a type of rating scale used in Job Task Analysis (JTA) surveys?
A. Likert Scale
B. Guttman Scale
C. Sullivan Scale
D. Stem Scale

Correct Answer: A. Likert Scale
Topic ID:2.2
(finish reason = stop; usage = 383 tokens)

Path: ./psy-003
Item 28:
What rating scale should be used for surveys when gathering job task analysis (JTA) data?
A. 7-point Likert Scale
B. 4-point Likert Scale
C. 10-point Likert Scale
D. 5-point Likert Scale

Correct Answer: D. 5-point Likert Scale
Topic ID:2.2
(finish reason = stop; usage = 398 tokens)

Path: ./psy-003
Item 29:

Scenario: You are working on data management for an educational assessment. You need to score the likert items in the data set. Which of the following methods would be most appropriate?

A) Summed Score
B) Item-Level Score
C) Scale Score
D) Mean Score

Correct Answer: A) Summed Score
Topic ID:3.1
(finish reason = stop; usage = 389 tokens)

Path: ./psy-003
Item 30:

Scenario: You are a Psychometrics who needs to score Likert items for a reliability analysis. How should you approach this?
A) Convert the Likert items into binary items
B) Compute the mean of the Likert item responses
C) Compute the median of the Likert item responses
D) Compute the mode of the Likert item responses

Correct Answer: B) Compute the mean of the Likert item responses
Topic ID:3.1
(finish reason = stop; usage = 426 tokens)

Path: ./psy-003
Item 31:

Scenario: You are developing a multiple-choice knowledge-based exam and you have omitted items with missing data. What is the best way to treat the missing data?

A. Set the missing data values to zero

B. Ignore the missing data
C. Assign random values to the missing data
D. Calculate the mean of the data and assign it to the missing values

Correct Answer: D. Calculate the mean of the data and assign it to the missing values
Topic ID:3.1
(finish reason = stop; usage = 428 tokens)

Path: ./psy-003
Item 32:

Scenario: When scoring items on a multiple-choice knowledge-based exam, how should missing data for omitted items be scored?

A. By providing no points
B. By providing a point for a correct answer
C. By providing points based on the difficulty of the item
D. By providing a proportion of the total points for the item

Correct Answer: A. By providing no points
Topic ID:3.1
(finish reason = stop; usage = 419 tokens)

Path: ./psy-003
Item 33:

Scenario: You are creating a knowledge exam for a psychometrics position. If the candidate runs out of time, how should the remaining items be scored?

A. Score the remaining items as incorrect
B. Score the remaining items as correct
C. Score the remaining items based on the percentage of items completed
D. Score the remaining items as half correct and half incorrect

Correct Answer: A. Score the remaining items as incorrect
Topic ID:3.1
(finish reason = stop; usage = 427 tokens)

Path: ./psy-003
Item 34:

Scenario: You are administering a knowledge-based multiple-choice exam and need to score not-reached item responses. What is the correct action to take?

A. Assign a score of zero to all not-reached items.
B. Assign a score of one to all not-reached items.
C. Assign a score of one-half to all not-reached items.
D. Assign a score of one-third to all not-reached items.

Correct Answer: A. Assign a score of zero to all not-reached items.
Topic ID:3.1
(finish reason = stop; usage = 457 tokens)

Path: ./psy-003
Item 35:

Scenario: You are tasked with cleaning exam data for analysis. What are the steps for cleaning the data?
A. Assemble, order, recode, join, subset, and store datasets
B. Wrangle data, analyze, clean, and store datasets
C. Analyze data, clean, recode, join, subset, and store datasets
D. Wrangle data, assemble, order, recode, join, subset, and store datasets

Correct Answer: D. Wrangle data, assemble, order, recode, join, subset, and store datasets
Topic ID:3.1
(finish reason = stop; usage = 438 tokens)

Path: ./psy-003
Item 36:

Scenario: Two variables, x and y, have been collected with the intention of evaluating their relationship. What is the best method for detecting bivariate outliers?

A. Calculating the linear correlation coefficient
B. Using a boxplot to identify outliers
C. Computing a t-test to identify outliers
D. Using the Mahalanobis distance

Correct Answer: D. Using the Mahalanobis distance
Topic ID:3.1
(finish reason = stop; usage = 413 tokens)

Path: ./psy-003
Item 37:

Scenario: You are a Psychometrician and have been asked to review examination data for quality control. What criteria should you use to reject cases from the data?

A. Data that is incomplete or inconsistent
B. Data that violates the assumptions of the analysis
C. Data that is not representative of the target population
D. Data that is too difficult to accurately measure

Correct Answer: A. Data that is incomplete or inconsistent
Topic ID:3.2
(finish reason = stop; usage = 417 tokens)

Path: ./psy-003
Item 38:

Scenario: You are conducting an examination data analysis and must decide on criteria for rejecting cases. What criteria would you use?
A. Ensure that the cases are not outliers
B. Confirm that all cases meet the criteria of the research hypothesis
C. Make sure that all cases have complete data
D. All of the above

Correct Answer: C. Make sure that all cases have complete data
Topic ID:3.2
(finish reason = stop; usage = 412 tokens)

Path: ./psy-003
Item 39:

Which of the following techniques is the most effective in detecting cheating on exams with correct and incorrect responses?
A) Exploratory data analysis
B) Item analysis
C) Response analysis
D) Item clustering

Correct Answer: C) Response analysis
Topic ID:3.2
(finish reason = stop; usage = 382 tokens)

Path: ./psy-003
Item 40:

Scenario: During a Psychometrician's analysis of a student's responses on an exam, they come across an unexpected number of correct and incorrect responses. What would be the best way to detect cheating on an exam with correct and incorrect responses?

A. Compare the student's response to the most common response on the exam
B. Calculate the student's score and compare it to the average score of other students
C. Examine the pattern of correct and incorrect responses
D. Analyze the student's demographic information

Correct Answer: C. Examine the pattern of correct and incorrect responses
Topic ID:3.2
(finish reason = stop; usage = 462 tokens)

Path: ./psy-003
Item 41:

Scenario: A psychometrician is presented with a test that has a reliability coefficient of 0.80. What should they do to account for attenuation?

A) Increase the sample size
B) Replace the items

C) Increase the difficulty level
D) Multiply the reliability coefficient by the square root of two

Correct Answer: D) Multiply the reliability coefficient by the square root of two
Topic ID:6.1
(finish reason = stop; usage = 399 tokens)

Path: ./psy-003
Item 42:

Scenario: You are a Psychometrician and are responsible for creating and maintaining exam blueprints. You have been asked to explain the strengths and weaknesses of methods for estimating psychometric reliability.

A. The Spearman-Brown prophecy formula is an accurate and reliable method to measure psychometric reliability.
B. The Cronbach's alpha coefficient is a useful method to measure psychometric reliability, but it has limited applications.
C. The Kuder-Richardson Formula is a powerful and reliable method to measure psychometric reliability.
D. The Rasch Model is an effective method to measure psychometric reliability, but it has some limitations.

Correct Answer: B. The Cronbach's alpha coefficient is a useful method to measure psychometric reliability, but it has limited applications.
Topic ID:6.1
(finish reason = stop; usage = 481 tokens)

Path: ./psy-003
Item 43:

Scenario: You are working on a psychometric project and need to calculate the conditional standard error for an item response theory (IRT) model. Which of the following is the correct formula for the conditional standard error?

A. SEconditional = SEitem / SEmodel
B. SEconditional = SEmodel / SEitem
C. SEconditional = SEitem - SEmodel
D. SEconditional = SEmodel - SEitem

Correct Answer: B. SEconditional = SEmodel / SEitem
Topic ID:6.1
(finish reason = stop; usage = 421 tokens)

Path: ./psy-003
Item 44:

Scenario: A psychometrician has just completed a multiple-choice exam and is analyzing the results. The psychometrician needs to determine if the exam items measure a single trait. What is the best psychometric technique to use?
A. Factor analysis
B. Item Response Theory
C. Classical Test Theory
D. Correlation analysis

Correct Answer: B. Item Response Theory
Topic ID:6.1
(finish reason = stop; usage = 398 tokens)

Path: ./psy-003
Item 45:

Scenario: A psychometrician finds that the responses of a group of examinees to a new test appear to fit a unidimensional model. To further test the assumption of unidimensionality, which of the following techniques should the psychometrician use?

A. Item response theory
B. Exploratory factor analysis
C. Item mapping
D. Item response curves

Correct Answer: B. Exploratory factor analysis
Topic ID:6.1
(finish reason = stop; usage = 410 tokens)