The creative psychometric item generator: a framework for item generation and validation using large language models

Antonio Laverghetta Jr.[1], Simone Luchini[1], Averie Linnell[2], Roni Reiter-Palmon[2] and Roger Beaty[1]

[1]*Department of Psychology, The Pennsylvania State University, 201 Old Main, University Park, Pennsylvania, USA*
[2]*Department of Psychology, University of Nebraska at Omaha, 6001 Dodge Street, Omaha, Nebraska, USA*

Abstract

Increasingly, large language models (LLMs) are being used to automate workplace processes requiring a high degree of creativity. While much prior work has examined the creativity of LLMs, there has been little research on whether they can generate valid creativity assessments for humans despite the increasingly central role of creativity in modern economies. We develop a psychometrically inspired framework for creating test items (questions) for a classic free-response creativity test: the creative problem-solving (CPS) task. Our framework, the creative psychometric item generator (CPIG), uses a mixture of LLM-based item generators and evaluators to iteratively develop new prompts for writing CPS items, such that items from later iterations will elicit more creative responses from test takers. We find strong empirical evidence that CPIG generates valid and reliable items and that this effect is not attributable to known biases in the evaluation process. Our findings have implications for employing LLMs to automatically generate valid and reliable creativity tests for humans and AI.

Keywords

automated item generation, prompt engineering, artificial intelligence

1. Introduction
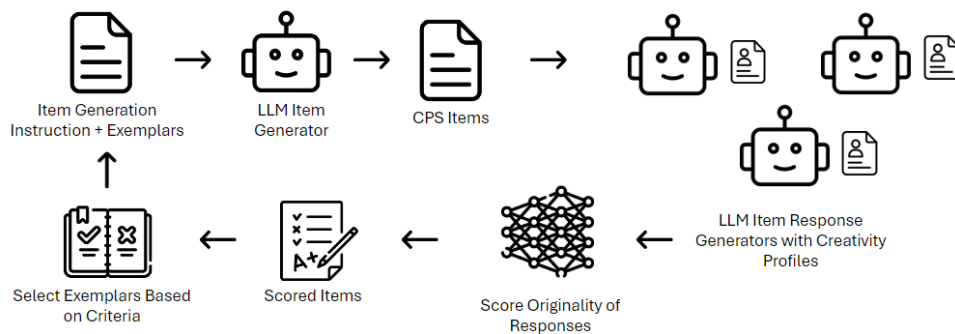
Figure 1: Overview of CPIG. From a base instruction, we prompt an LLM to generate CPS items, which are, in turn, completed by other LLMs. We give each LLM response generator a distinct profile to increase variability in the originality of their solutions. These responses are scored with an originality model developed by [1], and a subset of the generated items with highly original responses are selected to include in the prompt for the next round of item generation. This figure was designed using images from Flaticon.com.

Creativity is considered one of the primary factors that determine individual [2] and organizational [3] success in the modern economy. This is due to improved automation of routine tasks [4], the increasing complexity and ambiguity of problems organizations face, and projected growth of the

creative sectors of the economy [5]. As such, the development of validated creativity tests has become increasingly important. Nevertheless, generating new creativity assessments remains a resource-intensive process requiring many hours of trial and error to develop suitable items (questions). Such items can be highly complex, requiring participants to reason about intricate scenarios or design solutions to ambiguous problems [1], and therefore are difficult for even subject matter experts to develop.

With the introduction of modern large language models (LLMs) [6, 7] the ability of AI to automatically develop novel creativity tests appears increasingly plausible [8], and LLMs are already being used to automatically generate items measuring a variety of cognitive skills [9, 10, 11]. Applying similar ideas in creativity assessment could provide a method to generate valid and reliable creativity tests at scale, which would be beneficial for assessing creativity in both humans and AI. However, doing so may also be contentious for some, given the broader debate on whether AI can be creative. Despite some evidence pointing towards AI creativity, whether AI-generated ideas are truly novel remains a hotly debated topic [12, 13]. Some research suggests that using LLMs may lower the diversity of ideas produced over time, resulting in reduced collective novelty [14, 15]. Public perception of the creativity of AI also remains mixed; humans tend to view creative works produced by AI as less novel than those produced by other humans [14], and this could be problematic if humans become aware that they are being given AI-generated creativity tests. Broader research in social psychology has found that LLMs produce highly similar responses to questions regarding political orientation, moral philosophy, and other complex constructs that usually exhibit high variability in humans [16]. Collectively, these results point to a diminished diversity of thought in LLMs, which has important implications for whether and how LLMs should be used to automate creativity assessment.

How can we employ LLMs in designing items for measuring creativity without comprising the validity of any conclusions drawn from such items? We approach this from a *psychometric* perspective, which is both a field dedicated to measuring psychological constructs in humans and the source of a rich body of work measuring similar constructs in AI [17, 18, 19]. When measuring a construct like creativity, psychometrics requires that any measurement be both valid and reliable — it must accurately measure the intended construct and give consistent results over repeated measurements. Accomplishing this involves developing tests whose items accurately measure the construct, which historically was done by human experts. Can we use LLMs to generate high-quality items for measuring creativity? If so, this would be invaluable not only for the study of human creativity but it might also allow us to measure creativity more accurately in LLMs, which would be a boon for assessing AI creativity. Nevertheless, no prior work has investigated whether LLMs can automatically generate creativity assessments.

In this paper, we develop a framework to extend item generation into the creativity domain: the *creative psychometric item generator* (CPIG). CPIG relies on structured prompting and psychometrically based exemplar selection to generate items for a creative problem-solving task (CPS), an influential test of creativity [20]. Our framework is iterative and allows us to continuously refine the same item based on automated validity metrics until reaching a desired level of quality. While other works have explored how to use LLMs to solve [21] and generate [22] CPS-like items, none to our knowledge has examined how to generate psychometrically rigorous assessments of creativity. We find that CPIG generated items are just as valid and reliable as those written by humans. Remarkably, LLM solutions to CPIG items also appear to become more original over successive rounds of generation, suggesting a possible method to boost the creativity of generative AI via carefully designed items.

We make the following contributions:

1. We develop CPIG, a new framework for generating creativity items using LLMs.[1]

2. Through a series of experiments, we confirm that CPIG generated items are just as valid as those written by humans, and that our metrics for validity are robust to known biases in the scoring process.

---

[1]Code and supplementary materials will be provided at: https://osf.io/umnk5/

2. Background

Creativity is thought to comprise multiple facets, including originality (the novelty of an idea) and effectiveness (how useful or relevant the idea is), among others [23]. Past work has demonstrated that human judgments of originality are an effective predictor of the creativity of ideas [23]. As such, the value of a creativity test rests on its capacity to elicit many original responses [24]. To measure originality, researchers historically relied on human judgments performed by trained raters — a method called the Consensual Assessment Technique (CAT) [25]. In the CAT, human raters are instructed to read a series of ideas and assess their originality on a Likert scale. Although effective, human scoring is not efficient, as the recruitment and training of human raters is often costly and prone to errors. More recently, automated creativity assessment tools have been developed, including finetuning LLMs to predict human creativity ratings [1]. Highly accurate models have been reported, often matching or surpassing the agreement between human raters, which makes it practical to evaluate the quality of creative responses at scale.

From a psychometric perspective, measuring an individual's creativity requires developing structured tasks to evaluate how well they can produce ideas that are both original and high quality. We focus on a CPS as the basis for our experiments. In this task, a participant is given a scenario involving a dilemma to be solved (e.g., a coworker's roommate is causing problems at work, and it may put both of their jobs at risk), and they must produce a creative solution to this dilemma [1]. Scenarios are ambiguous by design, with many possible solutions, and reflect creative thinking in day-to-day settings. We focus on this CPS task due to its popularity as a creativity test and the availability of automated and psychometrically validated models for assessing the originality of CPS responses [1]. However, because many creative tasks can be evaluated in terms of originality, our methods are extensible to other tasks that can be automatically scored.

3. The architecture of CPIG

We take a psychometric approach to generating CPS items, inspired by recent work on automatically generating psychometrically valid test items [11, 9, 17]. We use LLMs to act as *item generators* to write the items, *item response generators* to create human-like solutions to the items, and *item scorers* to score the originality of LLM responses using psychometrically validated metrics. We hypothesize that originality in item responses provides a proxy for item quality: items with high quality should enable more creative responses and will tend to elicit better originality scores on average than those that are of lower quality. Optimizing for originality thus provides a way to generate higher quality items that can better tap the creative potential of subjects. Figure 1 shows an overview of CPIG.

3.1. Item generation

Automatically generating valid CPS items is a non-trivial task, as the items must describe sufficently complex scenarios to allow a wide variety of responses while also being sufficiently ambiguous that no single solution is canonically more "correct" than the others. Furthermore, we also want scenarios to describe a wide range of situations to avoid generating an item pool revolving around a narrow range of topics. We thus develop a multi-stage prompting method.[2]

First, before any runs of CPIG, we first prompt GPT-3.5-TURBO to generate lists of words, where each list contains three names, a place, and an action (e.g., "Mark", "beach", "Amy", "Lucas", "swimming"). The goal behind this step is to make the item generation task more concrete; rather than prompting the item generator LLMs to design scenarios without any additional context, we instead use the word lists as criteria that must be satisfied (e.g., the final scenario must contain all the names from the word list). This is meant to both simplify generation by breaking it down into multiple steps and help maximize diversity in scenario content by using different word lists to ensure no two item generation prompts

---

[2]All prompts used throughout CPIG are listed in the supplementary material.

are the same. We have GPT-3.5-TURBO generate ten word lists at once to help eliminate redundant lists and query the model five times to generate 50 lists in total. We set the max number of tokens to 2048 and the temperature to 1.0, leaving other parameters at their defaults. We use this process to generate lists covering a wide variety of semantic content that we manually checked to confirm they obeyed the specified format. We use these word lists throughout all trials of CPIG.

We use these word lists in the item generation prompt, where we instruct item generator LLMs to design CPS items using the contents of the word list provided. We provide LLMs with generation guidelines and examples of CPS items written by experts. For each trial, we attempt to generate one scenario for each word list. However, the generated items may fail basic validity checks for a variety of reasons, so to mitigate this, we develop a list of rules to drop generations that are likely low quality:

1. We compute item readability using Flesch's reading ease [26] and drop scenarios with scores lower than 45 (considered very difficult to read). We note that this metric requires a minimum string length to compute, so we also require that scenarios be at least 140 tokens long. We use the NLTK word tokenizer to ensure a conistent token count.[3]

2. From preliminary trials, we find that LLMs sometimes generate scenarios with priming effects, steering participants toward specific solutions. Examples of this include generating a list of possible solutions or setting up the scenario as a dichotomy ("Should I do X or Y?"). Based on the content of such scenarios, we developed a list of strings that indicate possible priming and drop scenarios that contain any such string. Specifically, we drop scenarios containing "on the one hand," "on the other hand," "dilemma," "must navigate," "must decide," "has to decide," and "is torn between." We do not claim that this list is comprehensive, but we found that it eliminated most priming in generated scenarios.

3. To prevent LLMs from generating irrelevant content after the scenario, we instruct them to always generate "I am finished with this scenario." at the end. We drop scenarios that lack this string.

Importantly, our goal behind this quality control was not to identify every possible error that might occur in the items, as we expect human experts will make the final decision for which items to include in a creativity assessment [9]. Rather, we use it to reduce the number of items that need to be examined by eliminating those that are unlikely to be valid. We attempt to generate a scenario a maximum of 10 times for each word list and drop the list if the LLM fails to generate a valid scenario on all attempts. We strip extra newlines and whitespace surrounding the scenario and text after the termination string (including the string itself).

3.2. Item response generation

Once we have LLM-generated items, we must evaluate whether they elicit creative responses. LLMs have proven adept at modeling psychometric data [19] and are competent as human simulacra for sociological modeling [27], so we use LLMs to generate synthetic responses to each item. A potential challenge here is that the item response generator LLMs may suggest similar solutions to the same item [14]. We account for this by adopting several prompting styles meant to increase the variation in the LLM responses: a *baseline* prompt where the LLM is asked to provide a creative solution to the item (with no further context), a *demographic* prompt where the LLM is provided demographic data about a hypothetical participant that it is meant to simulate while responding (e.g., "You are a Hispanic woman who works in real estate"), and a *psychometric* prompt where we replace the prior demographic data with statements sourced from psychometric inventories strongly correlated with creative performance.

For demographic and psychometric prompts, we construct a pool of *participant creativity profiles* to draw from based on responses to prior creativity studies [1]. These responses include differing occupations and responses to psychometric assessments, which we reason would increase the variability

---

[3]https://www.nltk.org/api/nltk.tokenize.word_tokenize

in the output of the item response generator LLMs. We provide demographic data in the prompt using either a variable format (e.g., "You are an Asian man") or as demographically relevant names. Demographic variables, including name, ethnicity, and gender, were taken from the New York City Health Department 2016 census of baby names,[4] and last names specifically were taken from the Decennial Census Survey[5] from the United States Census Bureau. We selected the three most common first and last names associated with each demographic variable for a total of 20 first names and 20 last names. We extract data for the psychometric prompts from a series of validated scales measuring constructs related to creativity. We employed scales tapping creative self-efficacy [28], creativity anxiety [29], creative mindset [30], openness to experience [31], tolerance for ambiguity [32], cynicism [33], and the RIASEC interest types [34].

In each prompting style, the model is provided a CPS item after the task instructions and demographic/psychometric profile (if applicable), and we process the generated response by removing extra newlines and white space. Because response generation is comparatively a much simpler task than item generation, we do not include additional content validity checks. We generate between 10 to 20 responses for each item. For the demographic and psychometric prompts, we sample a participant profile at random each time.

3.3. Item scoring and selection

Each LLM-generated item response is then scored using the methodology developed by [1], which trained ROBERTA-BASE [35] to predict mean originality scores of responses to CPS items. Specifically, this model was trained on a dataset annotated by experts for originality, who scored each response using a five-point Likert scale. They used a test set comprising originality scores to CPS items not seen during training and obtained a 0.41 Pearson correlation with human ratings. We use this model to score the originality of each CPIG item, which we use to select k items to include as exemplars in the next round of item generation. We develop several shot selection strategies for choosing exemplars, which we discuss below. Additionally, we include a baseline that simply chooses k items at random.

3.3.1. Greedy

This approach simply selects the k items with the highest originality scores. Specifically, we take the mean of the originality scores of all the responses per item and sort the resulting scores to select the k items with the highest scores.

3.3.2. Constraint satisfaction

A challenge with the greedy approach is that it may choose highly similar items if they all score high on originality. Indeed, we found in preliminary trials that cosine similarity scores between all pairs of the k items tend to increase over iterations, sometimes drastically. To address this, we develop another shot selection method that instead finds a set of k items that maximize originality and minimize similarity, which we treat as a constraint satisfaction problem. For each iteration of CPIG, we have a set of exemplars I from the prior iteration[6] with a mean originality score I_o and a mean semantic similarity I_v (the mean cosine similarity scores between all pairs of items in I). Additionally, we include thresholds δ_o and δ_v that define a tolerance above I_v and below I_o for the new set of exemplars. We then search for a set η of size k from the generated item pool at the current iteration that satisfies:

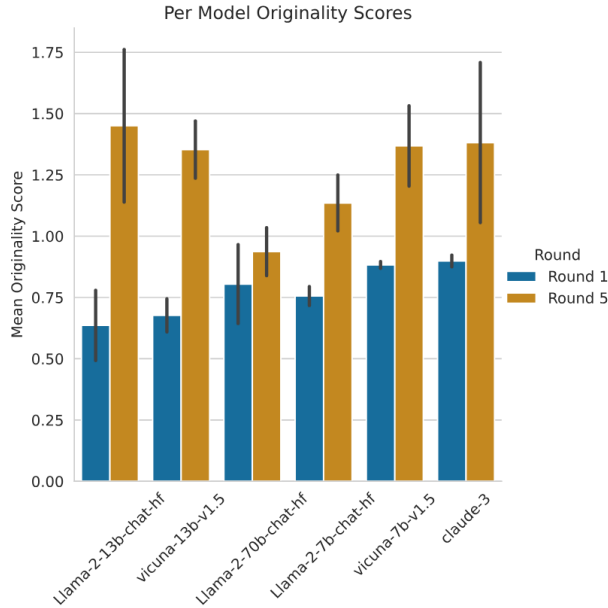$$\eta_o > I_o \ \lor \ I_o - \eta_o \le \delta_o \tag{1}$$

$$\eta_v < I_v \ \lor \ \eta_v - I_v \le \delta_v \tag{2}$$

We use Sentence Transformers [36] and ALL-MINILM-L6-v2 to compute I_v and η_v, and we search for all matching η across all unique combinations of size k from the item pool. We return the η with the highest originality score; further details on this method and the chosen values for δ are provided in the supplementary material.

Figure 2: Mean originality scores from each item generator on the first and last rounds, for all trials that did not use random shot selection. Error bars are standard deviations in scores. Higher values indicate more original item responses, on average.

3.4. Implementation details

We implement CPIG using LangChain[7] and utilize a variety of chat-based open-source and commercial LLMs, including LLAMA-2 (7b, 13b, and 70b) [37], VICUNA-1.5 (7b and 13b) [38], and CLAUDE-3-HAIKU.[8] All open-source models are implemented using Transformers [39]. We set the temperature to 1.0 across all trials to increase variation in the generated items and responses while leaving other text generation parameters at their defaults. We select four items to use as exemplars for all shot selection methods to ensure item generation prompts do not become too long and because we find this is sufficient to ensure variation in item content. We cap item generation to a maximum of 768 tokens and item response generation to 350 tokens, as responses to CPS items tend to be much shorter than the items themselves. We run each CPIG trial for five iterations, using three random seeds for every hyperparameter combination. We use the same LLM for item generation and item response generation for each open-source model trial and use LLAMA-7B for response generation when using CLAUDE-3-HAIKU for item generation. We provide a table listing all trials in the supplementary materials. We run experiments on three Nvidia RTX A6000 GPUs with 49GB of video memory each. We apply 4-bit quantization to all supported models.

4. Results

We present a comprehensive picture of how effective the different components of CPIG are at generating items that maximize the originality of the output from item response generator LLMs. This includes both ablations on the effect of the different prompting strategies and shot selection methods,

---

[7]https://www.langchain.com/
[8]https://www.anthropic.com/news/claude-3-family
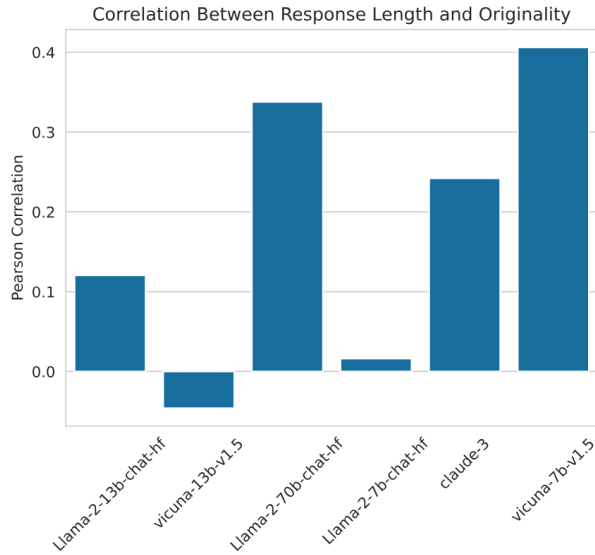
Figure 3: Pearson correlation between item response length and originality score. Length is calculated using the NLTK word tokenizer.

as well as human review on the quality of the generated items. For any ablation that requires computing semantic similarity, we use Sentence Transformers [36] and ᴀʟʟ-MɪɴɪLM-L6-v2 as the embedding model. All density plots employ kernel density estimation [40].
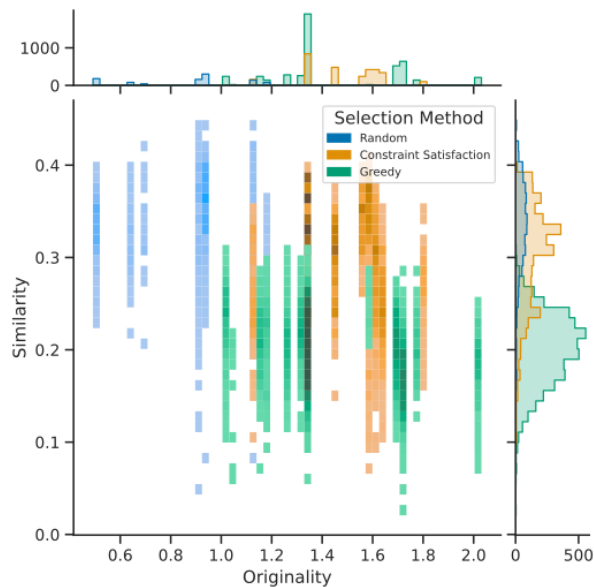

Figure 4: Joint histogram of originality and similarity scores for round five items. The highest quality items are those in the bottom right region. Note that we have dropped all items whose cosine similarity was greater than 0.95 to any other item.

4.1. Originality of LLM responses

Figure 2 shows originality scores for all runs that do not use random shot selection, broken down by model type. Critically, regardless of the item generator, CPIG consistently improves originality scores of responses by the last round of item generation, in some cases *more than doubling* the score compared to the first round. The difference in mean scores was significant in t-tests for both demographic ($p \ll 0.001$) and psychometric ($p \ll 0.001$) prompting styles and hence remains regardless of the specific prompting strategy used for item response generation. This demonstrates that CPIG-generated

items can elicit more creative responses from the item response generator LLMs. However, a potential confound when scoring originality is that the metric is influenced by the length of the response, with longer solutions typically being scored as more original [1]. We find that LLM responses are, on average, much longer than those of humans, leaving open the possibility that the increase in originality is driven purely by more elaboration in the response. We check for this by computing the Pearson correlation between response length and originality for every generation model and the items generated on the last round (not including random shot selection). Results are shown in Figure 3. As expected, length is at least partially correlated with originality for all generation models, though there is significant variation in the strength of this correlation. Importantly, however, the correlations remain weak overall and do not rise above 0.3 in either direction for most LLMs, suggesting that the increases in originality are not only due to increasing response length.
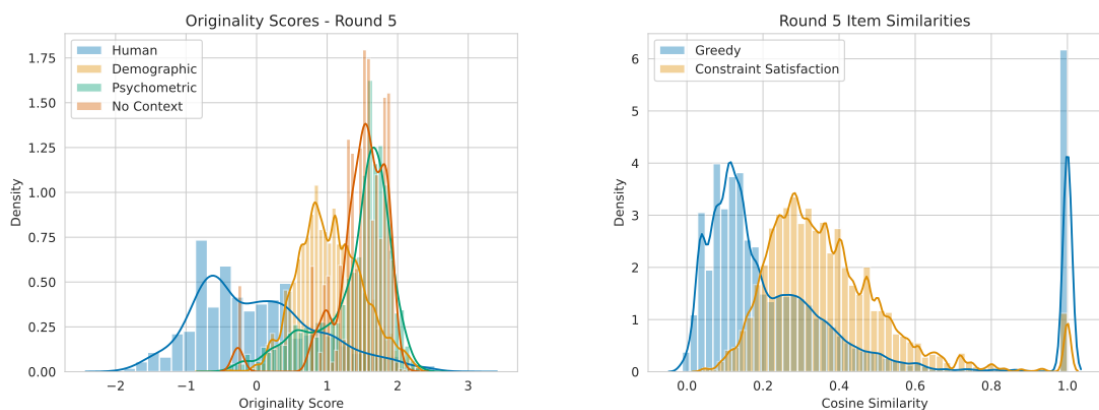
(a) Distributions of originality scores, broken down by item response prompting strategy. As a point of comparison, we also plot the originality scores of the human participants used to train the scoring model from [1], but note that they are not given the same items generated by CPIG.

(b) Cosine similarity scores between all pairs of items from the last round of generation, for both greedy shot selection and constraint satisfaction.

Figure 5: Distributions of originality (a) and similarity (b) scores, broken down by prompt types and shot selection strategy.

4.2. Relationship between originality and similarity

While improvements in response originality denote an increase in item quality, it remains unclear whether the item generator LLMs converge onto a few similar yet high-quality scenarios or how these variables relate to each other in the generated item pool. We explore this by plotting a joint histogram of originality and similarity scores[9] for all generated items, broken down by shot selection method, in Figure 4. Darker cells in this figure indicate a higher frequency of a particular originality-similarity combination. We observe that random shot selection obtains the worst combination of results: not only are most items low on originality, but the distribution also peaks the highest on similarity. Both greedy shot selection and constraint satisfaction achieve lower similarity and higher originality and do so consistently. As the originality of items produced using these strategies increases, their similarity scores remain generally static, indicating that improvements in originality do not come at the expense of more redundant items.

One notable trend is that greedy shot selection seems to have lower similarity scores on average despite constraint satisfaction being designed to minimize similarity. However, for this figure, we dropped all items whose similarity is above 0.95 to any other item to make computing the joint his-

---

[9]Measured as the mean cosine similarity between each item and every other item.
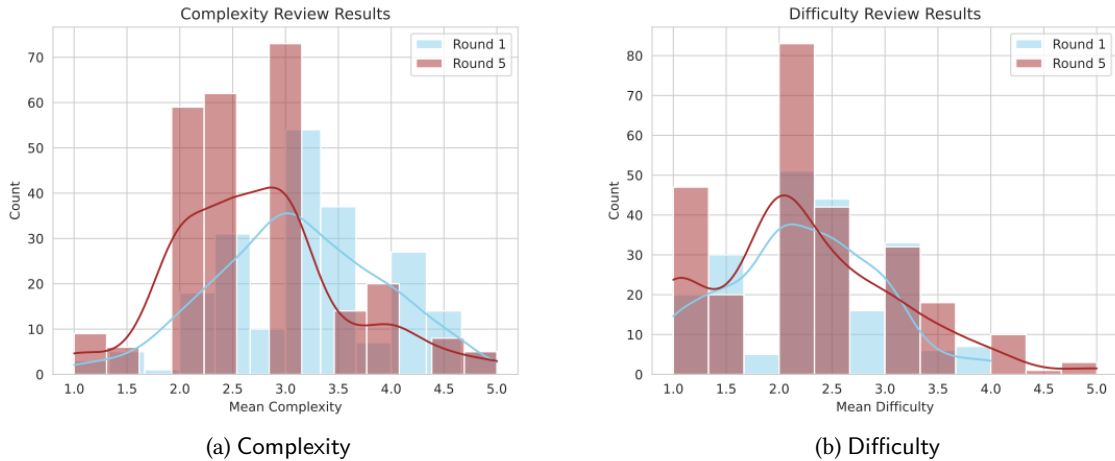
Figure 6: Mean complexity and difficulty scores from round one compared against round five. A rating of three indicates ideal complexity/difficulty.

togram more manageable. In Figure 5, we graph the univariate histogram of cosine similarity scores for both greedy and constraint satisfaction, and this time, include all the items that are generated in the last round. Although both methods generate some item pairs with cosine similarities of 1.0, there are many more such items for greedy shot selection, indicating a much larger fraction of extremely similar item content. Interestingly, greedy also peaks at a higher density than constraint satisfaction towards the lower end of the distribution. This likely reflects the balancing act required for constraint satisfaction; selecting items to maximize originality may sometimes require increases in similarity, though the method still succeeds in eliminating most duplicate content.

4.3. Effect of item response prompting style

Humans typically exhibit high variability in the originality of their responses to CPS items [1]. The different item response prompting strategies we develop are meant to induce a similar degree of variation, and we examine how effective they are in Figure 5. Compared to the no-context baseline — where the item response generator LLMs are simply instructed to answer the item — both demographic and psychometric prompting strategies exhibit higher variance and heavier tails in the originality distribution, better reflecting the trends from human participants. Both curves still have lower variance than humans and much higher peaks in originality scores, so it appears there remains headroom for alignment between LLM and human psychometric properties. The main challenge here again relates to elaboration in the response; while human participants often give short solutions, LLMs tend to provide very elaborate responses that embed multiple solutions simultaneously. Fully overcoming this challenge requires more sophisticated prompting and perhaps additional finetuning on human responses to align with our preferences for this task, but we leave this to future work.

4.4. Human content review

The prior results demonstrate that, with carefully chosen prompts and few-shot exemplars, `CPIG` can generate items that elicit more original responses from LLM test takers. But is this trend due to improvements in item quality or some other artifact of the generation process? We explore this by recruiting human annotators to rate the quality of the `CPIG` items.

We recruited five annotators with prior experience in rating for creativity studies. Annotators rated each item in terms of its *complexity* and *difficulty*, where we define complexity as how many *demands* were present in the item and difficulty as how many of those demands directly compete with each other, such that a solution that attempts to solve one might come at the expense of another. We define

demands as any relevant information in the scenario that could be used to construct a creative solution. Demands could include challenges to overcome in the scenario or resource constraints, among many others. We selected these facets to cover the most important factors to rate to ensure content validity in the items based on our expertise in creativity assessment and preliminary examinations of the items generated by CPIG. Both facets were rated on a five-point Likert scale, with one being too simple/easy, five being too complex/difficult, and three having the right amount of complexity/difficulty. This scale allowed us to account for both extremes of item content; items that are too complex or difficult might cause human participants to give up prematurely, while items that are too simplistic or easy are unlikely to require much creativity to solve. We designed a rubric that annotators used to rate each item, including definitions for complexity and difficulty. The annotators were first shown the rubric and allowed to ask any questions they had about the task. Then, together with one of the authors, the annotators rated ten practice items. Finally, the annotators, in combination with two of the authors, rated the remaining items via a missing data approach, where annotators only rated a subset of the CPIG items. This approach allowed us to achieve maximum coverage of all items while limiting rating time and making the annotation workload manageable. Each annotator rated between 200 and 245 LLM-written items, including items from the first and last round of CPIG. Annotators were only provided the text of each item, and were blinded to all other related details. For instance, annotators were not informed of which items belonged to which round of CPIG.

We obtained intraclass correlations of 0.52 for complexity and 0.49 for difficulty, for absolute agreement on the average ratings, indicating a modest rater agreement.[10] We plot in Figure 6 the distributions of complexity and difficulty scores from the items from the first and last rounds. For complexity, we see a definite improvement from round five, with a much larger fraction of items achieving the ideal complexity level than was present in round one. Trends are more static for difficulty as the distributions are quite similar to each other, especially at the ideal difficulty level. Collectively, the content review indicated that CPIG items are generally of high quality and that later iterations result in definite improvements for at least some facets of item quality.

We include two items generated by LLama-13b in Table 1, both using the same word list. While even items generated in the first round exhibit many desirable qualities, we see key improvements over iterations. Although the round one item (top row in the table) sets up what could be a complex scenario, it remains unclear what the exact problem is other than that Noah is being asked to do "extra work" for a customer. The round five scenario (bottom row) makes this clear: a new family is causing problems by stealing plants. This scenario also introduces added complexity by including new characters with interwoven relationships, hence adding more competing demands that need to be considered. The scenario is still not perfect as not all the information appears especially relevant, but overall, it does appear to be both more original and of higher quality.

5. Related work

5.1. Psychometric AI

Psychometric analysis of language models has seen growing interest in NLP research [11, 19, 41, 18, 42, 43]. Measurement models from psychometrics provide a strong test bed for evaluating language understanding in LLMs [18], making psychometrics a valuable tool for building better NLP test sets. However, LLMs are also valuable for modeling psychometric properties exhibited by humans on both cognitive [19] and non-cognitive [10] assessments, spurring interest in how LLMs might model human response data more broadly [44]. One rapidly growing research area is automated item generation, where LLMs are used to create new test items for standardized assessments with little or no human intervention [9, 11]. Several works have proposed frameworks similar to ours, where multiple LLMs are used to iteratively generate and evaluate new test items [45, 17]. However, this research has focused almost entirely on generating multiple-choice items, where the range of possible responses is inherently

---

[10]This was expected as rating creativity can be highly subjective, so it is challenging to achieve stronger rater agreement.

Table 1

Example items generated by LLAMA-13B from the first (top row) and last (bottom row) iterations of item generation, using the same word list.

CPS Item
Noah is a 25-year-old man who works as a gardener at a local gardening company. He is known for his excellent gardening skills and is always willing to help his customers find what they need. Recently, a new customer named Lily has started coming to the gardening company, and she has been causing problems for Noah. Lily is always asking Noah to do extra work for her, and she has been making him feel overwhelmed. Noah is starting to feel frustrated and is unsure of what to do. He does not want to lose his job, but he also does not want to work in an environment that is so tense and hostile. He is considering talking to his manager about the situation, but he is worried that it could backfire on him. He is unsure of what to do.
Noah is a 35-year-old man who lives in a small town with his wife Lily and their 5-year-old son James. Noah is a skilled gardener and spends most of his time tending to the plants in the community garden. Lily is a painter and spends most of her time in her studio, but she also helps out in the garden when she can. James loves spending time in the garden with his parents and is always eager to help out. Recently, a new family moved into town and they have been causing problems for Noah and Lily. The new family, the Smiths, have been stealing plants from the community garden and selling them at the local farmer's market. Noah and Lily are not sure what to do about the situation. They do not want to confront the Smiths directly, but they also do not want to lose their plants. They are considering asking James's teacher, Ms. Johnson, for help. Ms. Johnson is a kind and fair person, but she is also a close friend of the Smiths. Noah and Lily are not sure if Ms. Johnson will be willing to help them or if she will be biased towards the Smiths. They are also worried that if they do ask Ms. Johnson for help, it could cause problems for James in school. They are at a loss for what to do.

restricted. Additionally, the constructs targeted by such frameworks are either purely cognitive (with an objectively correct answer) or non-cognitive (open to interpretation based on individual differences). Creativity does not neatly fit into either mold: there is an aspect of "correctness" when judging CPS responses as the goal is to present a viable solution, yet how solutions are compared against each other in terms of originality is often open to rater interpretation [46]. Our work thus moves psychometric AI in a new direction to examine constructs outside the narrow scope explored in prior work.

5.2. Prompt engineering for psychometric assessment

An often-overlooked aspect of AI-based test development is prompt engineering: the process of developing prompts for LLMs that yield strong performance on the task of interest. Many studies rely on manual prompt tuning to adapt LLMs to a specific cognitive or psychometric task, which has allowed for the successful replication of many classic results from cognitive psychology [47] and has yielded high-quality items for various assessments [10]. A typical design pattern for such prompts is to use a format that aligns closely with how the actual task is presented to humans as if to simulate an experimental session [44]. However, greater care must be taken in the prompt design than might be necessary for other applications, as LLMs appear susceptible to more biases in task instructions than humans [48]. A starting point for addressing this could be to employ methods for prompt optimization, which have been widely successful in improving the performance of LLMs for NLP tasks [49]. These techniques, while powerful, typically rely on information-theoretic metrics for assessing prompt quality, often resulting in uninterruptible prompts [50]. A few works have explored how to create prompt optimization methods employing psychometrics as optimization targets by combining LLM item generators with discriminative models trained to predict item alignment with a target construct [45] or by incorporating standard metrics for reliability and validity to assess the quality of an LLM's generations [11, 17]. Even in these cases, the prompt itself usually remains static. CPIG provides a structured method for prompt mutation via the selection of exemplars that demonstrate evidence of validity on the task of interest.

6. Conclusion

We propose `CPIG`, a framework for generating creativity items using LLMs. By combining state-of-the-art models for response scoring with methods for item generation, we find that `CPIG` can generate items that improve the originality of LLM responses over time, which in turn points to increased creativity in their solutions. This trend is not attributable to known biases in the scoring model, and human raters find `CPIG` items to be high quality.

While our results are promising, our analysis also has limitations. In developing `CPIG`, we focused primarily on originality as the metric to optimize. While originality is a crucial facet of creativity, it is just one metric for judging creative outputs. Depending on the context, other metrics, such as an output's quality or relevance, may be more important to evaluate, and future work should extend our framework to optimize multiple criteria simultaneously. The quality of the generated items depends directly on the item evaluation, which was accomplished through automated scoring that, while effective, is not without limitations [1]. Developing more robust evaluations requires layering multiple quality control checks on top of each other, perhaps by employing separate LLM judges to rate the quality of the items directly and provide structured feedback on how to improve the items. Though we performed a content review on the `CPIG` items, it remains unclear how effective they would be when administered to human participants to solve without conducting more studies. As such, we caution against using the items from `CPIG` until they have undergone more extensive review. Finally, we must acknowledge biases in the LLMs, which may have influenced item generation. The data for our scoring model was curated using raters from a Western background [1], making the possibility of bias even more likely. Addressing this requires curating originality scores representing a more diverse slate of cultural views and developing bias mitigation strategies during item generation to ensure the evaluation remains fair.

Acknowledgments

References

[1] S. Luchini, N. T. Maliakkal, P. V. DiStefano, J. D. Patterson, R. Beaty, R. Reiter-Palmon, Automatic scoring of creative problem-solving with large language models: A comparison of originality and quality ratings (2023).

[2] C. Makó, M. Illéssy, Automation, creativity, and the future of work in europe: A comparison between the old and new member states with a special focus on hungary, MTA Társadalomtudományi Kutatóközpont Kisebbsegkutató Intézet (2020).

[3] W. Tsegaye, Q. Su, M. Malik, The antecedent impact of culture and economic growth on nationscreativity and innovation capability, Creativity Research Journal 31 (2019) 215–222.

[4] M. Chui, J. Manyika, M. Miremadi, Four fundamentals of workplace automation, McKinsey Quarterly 29 (2015) 1–9.

[5] T. M. Amabile, Creativity, artificial intelligence, and a world of surprises, Academy of Management Discoveries 6 (2020) 351–354.

[6] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al., On the opportunities and risks of foundation models, arXiv preprint arXiv:2108.07258 (2021).

[7] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam,

G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[8] J. Rafner, R. E. Beaty, J. C. Kaufman, T. Lubart, J. Sherson, Creativity in the age of generative ai, Nature Human Behaviour 7 (2023) 1836–1838.

[9] A. A. von Davier, A. Runge, Y. Park, Y. Attali, J. Church, G. LaFlair, The item factory, Machine Learning, Natural Language Processing, and Psychometrics (2024) 1.

[10] P. Lee, S. Fyffe, M. Son, Z. Jia, Z. Yao, 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 (2023) 163–190.

[11] A. Laverghetta Jr., J. Licato, Generating better items for cognitive assessments using large language models, in: E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, T. Zesch (Eds.), Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 414–428.

[12] S. Sæbø, H. Brovold, On the stochastics of human and artificial creativity, arXiv preprint arXiv:2403.06996 (2024).

[13] G. Franceschelli, M. Musolesi, On the creativity of large language models, arXiv preprint arXiv:2304.00008 (2023).

[14] B. R. Anderson, J. H. Shah, M. Kreminski, Homogenization effects of large language models on human creative ideation, arXiv preprint arXiv:2402.01536 (2024).

[15] A. R. Doshi, O. Hauser, Generative artificial intelligence enhances creativity, Available at SSRN (2023).

[16] P. S. Park, P. Schoenegger, C. Zhu, Diminished diversity-of-thought in a standard large language model, Behavior Research Methods (2024) 1–17.

[17] Y. Attali, A. Runge, G. T. LaFlair, K. Yancey, S. Goodwin, Y. Park, A. A. von Davier, The interactive reading task: Transformer-based automatic item generation, Frontiers in Artificial Intelligence 5 (2022) 903077.

[18] C. Vania, P. M. Htut, W. Huang, D. Mungra, R. Y. Pang, J. Phang, H. Liu, K. Cho, S. Bowman, Comparing test sets with item response theory, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 1141–1158.

[19] A. Laverghetta Jr, A. Nighojkar, J. Mirzakhalov, J. Licato, Can transformer language models predict psychometric properties?, in: Proceedings of* SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics, 2021, pp. 12–25.

[20] R. Reiter-Palmon, M. Y. Illies, L. Kobe Cross, C. Buboltz, T. Nimps, Creativity and domain specificity: The effect of task type on multiple indexes of creative problem-solving., Psychology of Aesthetics, Creativity, and the Arts 3 (2009) 73.

[21] S. R. Rick, G. Giacomelli, H. Wen, R. J. Laubacher, N. Taubenslag, J. L. Heyman, M. S. Knicker, Y. Jeddi, H. Maier, S. Dwyer, et al., Supermind ideator: Exploring generative ai to support creative problem-solving, arXiv preprint arXiv:2311.01937 (2023).

[22] Y. Tian, A. Ravichander, L. Qin, R. L. Bras, R. Marjieh, N. Peng, Y. Choi, T. L. Griffiths, F. Brahman, Macgyver: Are large language models creative problem solvers?, arXiv preprint arXiv:2311.09682 (2023).

[23] J. Diedrich, M. Benedek, E. Jauk, A. C. Neubauer, Are creative ideas novel and useful?, Psychology of Aesthetics, Creativity, and the Arts 9 (2015) 35.

[24] M. A. Runco, G. J. Jaeger, The standard definition of creativity, Creativity Research Journal 24 (2012) 92–96.

[25] P. J. Silvia, B. P. Winterstein, J. T. Willse, C. M. Barona, J. T. Cram, K. I. Hess, J. L. Martinez, C. A. Richard, Assessing creativity with divergent thinking tasks: exploring the reliability and validity of new subjective scoring methods., Psychology of Aesthetics, Creativity, and the Arts 2 (2008) 68.

[26] J. Kincaid, R. P. Fishburne Jr, R. L. Rogers, B. S. Chissom, N. T. T. C. M. T. R. Branch, Derivation of

new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel (1975).

[27] S. Sun, E. Lee, D. Nan, X. Zhao, W. Lee, B. J. Jansen, J. H. Kim, Random silicon sampling: Simulating human sub-population opinion using a large language model based on group-level demographic information, arXiv preprint arXiv:2402.18144 (2024).

[28] M. Karwowski, Did curiosity kill the cat? relationship between trait curiosity, creative self-efficacy and creative personal identity, Europe's Journal of Psychology 8 (2012) 547–558.

[29] R. J. Daker, R. A. Cortes, I. M. Lyons, A. E. Green, Creativity anxiety: Evidence for anxiety that is specific to creative thinking, from stem to the arts., Journal of Experimental Psychology: General 149 (2020) 42.

[30] M. Karwowski, Creative mindsets: Measurement, correlates, consequences., Psychology of Aesthetics, Creativity, and the Arts 8 (2014) 62.

[31] C. G. DeYoung, L. C. Quilty, J. B. Peterson, J. R. Gray, Openness to experience, intellect, and cognitive ability, Journal of personality assessment 96 (2014) 46–52.

[32] A. Furnham, T. Ribchester, Tolerance of ambiguity: A review of the concept, its measurement and applications, Current Psychology 14 (1995) 179–199.

[33] K. S. Mitchell, R. Reiter-Palmon, Malevolent creativity: personality, process, and the larger creativity field, in: Creativity and Morality, Elsevier, 2023, pp. 47–68.

[34] P. I. Armstrong, S. X. Day, J. P. McVay, J. Rounds, Holland's riasec model as an integrative framework for individual differences., Journal of Counseling Psychology 55 (2008) 1.

[35] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[36] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: http://arxiv.org/abs/1908.10084.

[37] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).

[38] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, E. P. Xing, Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023.

[39] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Q. Liu, D. Schlangen (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45.

[40] E. Parzen, On estimation of a probability density function and mode, The annals of mathematical statistics 33 (1962) 1065–1076.

[41] A. Laverghetta Jr, A. Nighojkar, J. Mirzakhalov, J. Licato, Predicting human psychometric properties using computational language models, in: The Annual Meeting of the Psychometric Society, Springer, 2021, pp. 151–169.

[42] Y. Li, Y. Huang, H. Wang, X. Zhang, J. Zou, L. Sun, Quantifying ai psychology: A psychometrics benchmark for large language models, arXiv preprint arXiv:2406.17675 (2024).

[43] J. He-Yueya, W. A. Ma, K. Gandhi, B. W. Domingue, E. Brunskill, N. D. Goodman, Psychometric alignment: Capturing human knowledge distributions via language models, arXiv preprint arXiv:2407.15645 (2024).

[44] M. Tavast, A. Kunnari, P. Hämäläinen, Language models can generate human-like self-reports of emotion, in: 27th International Conference on Intelligent User Interfaces, 2022, pp. 69–72.

[45] I. Hernandez, W. Nie, The ai-ip: Minimizing the guesswork of personality scale item development through artificial intelligence, Personnel Psychology 76 (2023) 1011–1035.

[46] M. Benedek, C. Mühlmann, E. Jauk, A. C. Neubauer, Assessment of divergent thinking by means

of the subjective top-scoring method: Effects of the number of top-ideas and time-on-task on reliability and validity., Psychology of Aesthetics, Creativity, and the Arts 7 (2013) 341.

[47] A. Ushio, L. Espinosa Anke, S. Schockaert, J. Camacho-Collados, BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies?, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 3609–3624.

[48] A. Gupta, X. Song, G. Anumanchipalli, Investigating the applicability of self-assessment tests for personality measurement of large language models, arXiv preprint arXiv:2309.08163 (2023).

[49] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, J. Ba, Large language models are human-level prompt engineers, arXiv preprint arXiv:2211.01910 (2022).

[50] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Computing Surveys 55 (2023) 1–35.