Original Research Article

Using Generative Artificial Intelligence to Advance Hypothesis-Driven Scale Validation: Identifying Criterion Measures and Generating Precise a Priori Hypotheses

Assessment
1–16

DOI: 10.1177/10731911251401321
journals.sagepub.com/home/asm
Sage

Kyle D. Austin[1], Hannah K. Crawley[1], William Fleeson[1], and R. Michael Furr[1]

Abstract

We propose, illustrate, and evaluate the use of artificial intelligence (AI) to advance rigorous hypothesis-driven scale validation. Using a qualitative approach, we found that AI provided useful suggestions for measures to be used as criteria in scale validation research. Using data and expert predictions previously used to validate nine scales/subscales, we evaluated AI's ability to produce precise, psychologically reasonable validity hypotheses. ChatGPT and Gemini produced hypotheses with "inter-trial consistency" similar to experts' "inter-rater consistency," and their hypotheses agreed strongly with experts' hypotheses. Importantly, their hypothesized validity correlations were roughly as accurate (in terms of corresponding with actual validity correlations) as were experts' hypotheses. Replicating across nine scales/subscales, results are encouraging regarding the use of AI to facilitate a precise hypothesis-driven approach to convergent and discriminant validity in a way that saves time with little-to-no cost in psychological or psychometric quality.

Keywords

validity, psychometrics, artificial intelligence, scale development, measurement, assessment

Introduction

Quality measurement is crucial for producing informative psychological research and for facilitating productive (and ethical) applied psychological work. Evaluating a measure's construct validity may be the most important facet of establishing the measure's quality, but doing so rigorously can be time-consuming and demanding. The purpose of the current work is to propose, illustrate, and evaluate the use of generative artificial intelligence (AI) to facilitate a rigorous hypothesis-driven approach to construct validation. Such use of AI may expedite and improve the validation process and thereby advance psychological assessment. Facilitating the validation process may increase researchers' ability to produce and evaluate new measures in a timely manner, and it may improve the quality of those measures

The Nature and Value of a Rigorous Hypothesis-Driven Approach to Scale Validation

Validity can be defined as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA et al., 2014, p. 11).[1] For example, when evaluating the validity of a scale intended to assess patience, researchers might gather evidence to test clear theoretically guided hypotheses regarding the meaning and correlates (and thus interpretation) of the scale's scores. Researchers would likely obtain the scale's correlations with one or more (sometimes as many as 10, 20, 30) criterion variables (e.g., Costello et al., 2022; Miller et al., 2016; Moultrie & Engel, 2017; Porcerelli et al., 2016; Poythress et al., 2010; Suzuki et al., 2017; Wehner et al., 2021). If those correlations are consistent with theory-driven hypotheses, then such consistency is taken as evidence of validity.

A rigorous approach to theory-testing involves confirmatory designs in which researchers specify precise a priori hypotheses and evaluate the degree to which data are consistent with those hypotheses. Such an approach is often adopted in some facets of psychometric evaluation (e.g., using confirmatory factor analysis to evaluate a scale's internal structure). However, a rigorous confirmatory approach also has strong potential value for evaluating

[1]Wake Forest University, Winston-Salem, NC, USA

Corresponding Author:
R. Michael Furr, Department of Psychology, Wake Forest University, 1834 Wake Forest Road, Winston-Salem, NC 27109, USA.
Email: FurrRM@wfu.edu

criterion validity, predictive validity, convergent validity, and discriminant validity. Those facets of validity can roughly be viewed in terms of whether a scale's scores *actually* correlate with other measures in ways that they *should* (based on the theoretical meaning of the relevant constructs) correlate with those measures. Researchers can evaluate those forms of validity evidence rigorously by identifying theoretically relevant criteria, generating precise a priori hypotheses regarding the correlations that should be obtained between a new scale and those criterion measures, and formally testing the degree to which data support or fail to support those hypotheses. Preregistration of such hypotheses, combined with quantitative evaluation of empirical support for those hypotheses, provides a highly rigorous and transparent approach to scale validation.

Such a rigorous hypothesis-driven approach to validation has enjoyed some use. For example, Reynolds et al. (2025) evaluated the convergent and discriminant validity of the Trait Truthful Communication Scale (T-TCS), which ostensibly assesses the tendency to engage in truthful communication. They chose 32 criterion measures roughly assumed to have varying relationships with truthful communication, and they generated a precise a priori hypothesis for each. For example, they hypothesized that a good measure of truthful communication should correlate $r = .58$ with the SAPA Personality Inventory (SPI) Honesty scale (reflecting convergent validity), should correlate $r = .10$ with the HEXACO-24 Openness scale (reflecting discriminant validity), and should correlate $r = -.48$ with the Pervasive Bullshitting scale. They then recruited participants to complete the T-TCS and all 32 criterion measures, computed the actual correlation between the T-TCS and each criterion, and found strong correspondence between hypothesized and actual correlations. For example, the T-TCS was correlated .61, .00, and −.45 with Honesty, Openness, and Pervasive Bullshitting, respectively. Thus, the criteria that were hypothesized to correlate positively, weakly, and negatively with a good measure of truthful communication did, in fact, correlate positively, weakly, and negatively with the T-TCS. This was interpreted as evidence that T-TCS scores could be validly interpreted as reflecting truthful communication.

Unfortunately, standard practice does not seem to use such precise a priori hypotheses (Wetzel & Killisch, 2025), particularly if more than a small number of criteria are examined. In evaluations of criterion validity, predictive validity, convergent validity, and discriminant validity, a standard approach is often to administer a new scale alongside one or more other measures, correlate the scale's scores with those measures, "eyeball" those correlations, and declare that the pattern of correlations is generally consistent with expectations. But those expectations are often not formally stated, and they might not even be considered until data are analyzed. Even less frequently are those expectations presented precisely (e.g., "we expect the correlation to be moderate to strongly positive" versus "we expect the correlation to be approximately .35"). This imprecision and/or lack of information leaves much up to subjective judgment, post hoc interpretation, and potential bias. Is a correlation of .25 too weak to be considered "moderate," or is a correlation of .50 or .60 stronger than *should* be obtained? This can weaken the rigor and quality of scale validation.

The Challenges of a Rigorous Hypothesis-Driven Approach to Scale Validation

There are at least two challenges limiting the use of a rigorous hypothesis-driven approach to scale validation. First, researchers might struggle to identify relevant criterion measures. Some constructs might appear to have few such criteria, or researchers might not be aware of measures available to assess a range of potentially relevant constructs. Even if researchers do generate ideas for criterion constructs and measures, the process can be lengthy and may be limited by the researchers' range of experience.

Second and usually more challenging, researchers might struggle to generate precise a priori predictions for one or more criterion measures. Even those with a clear theoretical understanding of a construct might be reluctant to commit to precise hypotheses regarding correlates of the construct, out of understandable concerns regarding their own idiosyncratic views or interpretations of the relevant constructs or of the strength of the potential correlations. One potential solution to this reluctance is to recruit topical experts to generate relevant hypotheses. A researcher might trust an expert panel's aggregate hypotheses more than her or his own potentially idiosyncratic predictions. Unfortunately, this solution imposes on those experts and may require significant time to execute. Experts, including colleagues, students, acquaintances, or others, are busy and likely have many commitments. Even experts eager to help may take several weeks to do so, delaying the validation process significantly and thus delaying development and evaluation of psychological measures. Moreover, the experts themselves might share biases or an inaccurate understanding of the relevant constructs.

AI Might Provide Solutions to Those Challenges

AI has exciting potential to enhance psychological assessment in diverse ways (e.g., Haines et al., 2025; Markey et al., 2025; Singh & Singh, 2024). Some researchers have explored its ability to produce scores based on open-ended responses, akin to structured or unstructured interviews (Dumas et al., 2025; Fan et al., 2023; Kjell et al., 2024). Others have evaluated its ability to produce scale items (Götz et al., 2024; Hernandez &

Nie, 2023; Krumm et al., 2024; Lee et al., 2023), with encouraging results.

We propose that AI can help address challenges that impede a rigorous hypothesis-driven approach to scale validation, and some recent work suggests that AI might well have this ability. Abdurahman et al. (2024) found that a trained transformer model could take an individual's responses to a set of items and predict, with reasonable accuracy, their responses to similar items. In fact, their model performed as well as human judges. Similarly, Banker et al. (2024) used AI, both GPT-3 and GPT-4, to produce hypotheses regarding social psychological phenomena. Expert raters evaluated these AI-produced hypotheses as being superior in multiple ways (e.g., plausibility, originality, and impact) to human-produced hypotheses (see also Tong et al., 2024). Thus, AI might be able to make psychologically reasonable predictions and hypotheses relevant to construct validation, and it can do so nearly instantly.

The Current Examinations

We examine AI's ability to facilitate rigorous hypothesis-driven validation. We first evaluate its ability to identify relevant criterion measures, as is necessary for many examinations of convergent and discriminant validity. AI might function as a useful colleague or research assistant by helping researchers consider or discover criterion measures that might otherwise remain unidentified. This evaluation will be qualitative and subjective.

Second, we evaluate AI's ability to generate precise hypotheses regarding validity correlations. We evaluate its ability to function as an expert, or panel of experts, in producing psychologically reasonable and precise hypotheses required in a rigorous hypothesis-driven approach to validation. If AI can provide such hypotheses, then researchers might significantly streamline or perhaps improve their validation work in ways that avoid imposing on (and waiting for) colleagues and acquaintances. AI might even provide hypotheses that are better or more accurate reflections of the relevant constructs than experts.

We use existing data from validation studies of nine scales/subscales. For each scale/subscale, experts provided precise hypotheses regarding correlations between a measure of each focal construct (e.g., truthful communication or patience) and criterion measures. Experts included faculty and graduate students with relevant theoretical expertise (e.g., in honesty, patience, moral psychology in general) and/or expertise in personality assessment. We will evaluate AI's performance as "an expert" in general and as compared to those original experts.

We do this for each scale/subscale using two AI models—ChatGPT 4o (available with paid subscription) and Google Gemini 2.0 Flash (available without cost). ChatGPT (developed by OpenAI) is, as of this writing, the most widely recognized and utilized large language model, making it the most likely candidate for use in academic tasks involving AI. For comparison and replicability, we also evaluate Google Gemini. This allows us to evaluate whether any poor (or strong) performance is unique to a single model. All correlations (hypothesized and actual) for all nine scales/subscales and all criteria measures, syntax for all analyses, and the online Supplemental document are available at https://osf.io/k9ptg/?view_only=1bd312e312f44e15830a03c4404a89b8

Can AI Help Identify Relevant Criterion Constructs and Measures?

To evaluate whether AI can facilitate the identification of relevant criterion constructs and measures, we adopted a simple qualitative approach. Our communication with AI and its responses are presented in our Online Supplemental Material. Using the T-TCS as a test case, we first provided ChatGPT 4o with a detailed description of truthful communication.[2] ChatGPT responded by summarizing its understanding of the construct. We then instructed:

> *Now, assume that I am carrying out the process of quantifying construct validity. Come up with 32 well-established scales that I could ask experts to predict correlations with that would give me a good idea of the convergent and discriminant validity of the construct.*

In response to this simple prompt, ChatGPT suggested 32 criterion scales. This list was organized in terms of convergent and discriminant validity and in terms of higher-order themes that ChatGPT extracted from the detailed description that we provided (e.g., "Moral and Ethical Behavior" and "Honesty and Integrity"). For each suggested scale, ChatGPT provided a citation and a brief summary of the relevant construct(s). For example:

> ***Moral Identity Scale*** (Aquino & Reed, 2002)*—Assesses the extent to which morality is central to one's self-concept.*

We found ChatGPT's suggestions to merit serious consideration. One suggestion (HEXACO-PI-R's Honesty/Humility scale) had been used by Reynolds et al. (2025), and many were similar to the criteria measures used by Reynolds et al. For example, ChatGPT suggested measures of the Big Five personality traits, intellectual humility, authenticity, moral tolerance, moral relativism, honesty, and deception, all of which were constructs assessed by Reynolds et al. In addition, several suggestions (e.g., measures related to guilt-proneness, curiosity, moral courage, and moral identity) were not reflected in Reynolds et al., but seem quite relevant and reasonable as considerations for convergent or discriminant validity criteria. Suggestions

reflected many measures likely to be positively correlated with a measure of truthful communication, some likely to be negatively associated with that construct, and some that might be weakly or uncorrelated. Overall, ChatGPT produced relevant suggestions that could have been used to facilitate the Reynolds et al.'s process of identifying constructs and measures for their scale validation work.

That said, several (8 out of the 32) of the suggestions were problematic in some way. Two scales did not seem to exist and thus were "dead ends" as suggestions. Other problematic suggestions were either mistitled or were accompanied by apparently incorrect citations (e.g., citing a paper that used a scale rather than the one in which the scale was originally presented), or both. These latter suggestions, though slightly problematic, did provide information that was potentially useful. In sum, although some of ChatGPT's suggestions required effort to identify and a small number were not useful, the majority were both clear and useful.

Can AI Generate Precise and Psychologically Plausible Validity Hypotheses?

To evaluate the quality of AI's hypotheses, we examined three issues. We first examined *consistency*, or the degree to which AI functions as "reliable" independent experts. When providing hypotheses multiple times, AI might produce highly inconsistent hypotheses. This is problematic within a set of experts (i.e., poor inter-rater reliability) and would be equally problematic with AI. We thus examine consistency within AI and compare it to consistency within panels of experts. We next examine *agreement*, or the degree to which AI's hypotheses converge with experts' hypotheses. If agreement is strong, then AI's hypotheses approximate those that would be provided by experts. Finally and most importantly, we examine *accuracy*, or the degree to which AI's predictions (regarding the validity correlations that should be obtained) correspond with validity correlations that are actually obtained. Even if AI agrees well with experts, it (and the experts) might be providing poorly grounded hypotheses. We will thus examine the degree to which AI's hypotheses are psychologically plausible and accurate.

Overview of Methods

Here we detail our evaluation using ChatGPT 4o,[3] with details of our analysis using Gemini in the Online Supplemental Material. We evaluated AI's ability to make psychologically reasonable predictions regarding convergent and discriminant validity, repeating this for nine focal scales/subscales. Each scale/subscale's validity has previously been examined through the hypothesis-driven procedure described earlier—gathering experts' predictions for validity correlations, obtaining data from participants and computing actual validity correlations, and examining correspondence between predicted and actual correlations.

The Moral Character Questionnaire (MCQ; Furr et al., 2022) comprises seven subscales—Global Morality, Loyalty, Honesty, Fairness, Compassion, Purity, and Respect. Each required its own psychometric evaluation, including convergent and discriminant validity. As reported in Furr et al. (2022), participants ($N = 246$ from Amazon's Mechanical Turk) completed the MCQ and 21 criterion measures.[4] The T-TCS (Reynolds et al., 2025) produces a score ostensibly reflecting "the tendency toward expressing one's beliefs accurately and faithfully." In an examination of the T-TCS's validity, $N = 525$ participants from Qualtrics Panels completed the T-TCS and 32 criterion measures (Reynolds et al., 2025). The Patient Reaction scale (P-React; Furr et al., in preparation) assesses the tendency to react patiently (i.e., the *tendency to react calmly, internally and externally, to slower than desired progress toward goal achievement*). To evaluate the P-React scale's convergent and discriminant validity, Furr et al. (in preparation) asked participants ($N = 242$ university students) to complete P-React and 22 criterion scales. Across the nine scale/subscale scores produced by these measures, we thus replicated our AI evaluation nine times for each chatbot (ChatGPT and Gemini).

For each evaluation, we provided AI with three pieces of information that were as identical as possible to information provided previously to experts (see the Online Supplemental Material for all instructions to both experts and AI). First, we provided instructions describing the hypothesis task. As shown in Figure 1 for our first test, we described the idea of a focal construct (e.g., Global Morality), the fact that we were asking for predictions, suggestions on how to make predictions, and a request to read the criterion scales' item content when making predictions (note: instructions for all tests are in the Online Supplemental Material). Second, we provided information about the focal construct or focal scale—either providing the construct name and scale items, or providing a conceptual definition/explanation about the construct without providing the focal scale's item content (again, reflecting what experts had received). Third, we provided information about each criterion scale—its name, item content (including indicating which items are positively keyed and which are negatively keyed), response scale, and a brief statement of the meaning of a high score on the scale (e.g., "a high score indicates greater filial piety"). In essence, we told AI (as experts had previously been told) what we wanted it to do, and we gave it the information necessary to do it. AI then produced a set of predicted correlations, each one a correlation between a scale measuring the focal construct and a criterion scale.

For each test, we repeated the process several times to match the number of experts previously providing predictions. As described below, we conceived of each AI "trial"

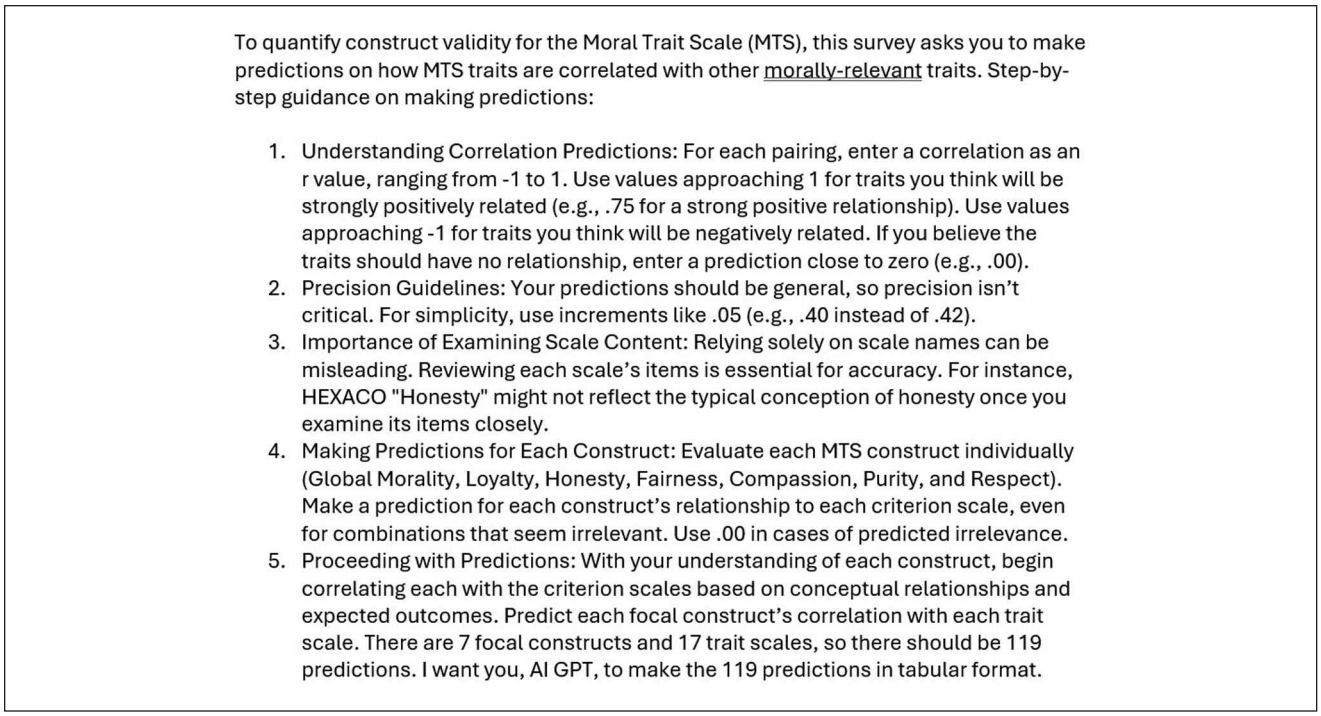

Figure 1. Instructions Given to ChatGPT, for the Moral Character Questionnaire.
Note. At an early phase of development, the MCQ was tentatively called the Moral Trait Scale. Since that term was presented to experts, we presented it to ChatGPT.

or set of predictions to parallel one independent expert. Thus, if five experts had previously provided predictions for a given focal scale, we asked AI to make its predictions five separate times. We disabled AI's memory function to ensure, as much as possible, that each AI trial is independent of previous trials (e.g., we disabled ChatGPT 4o's ability to remember previous chats). To some degree, this parallels work by independent experts.

Methodological Differences Across Focal Scales

The original validation work for the MCQ, T-TCS, and P-React differs from each other in two ways relevant to the current work. First, they differ regarding the availability of directly relevant online information at the time of our analysis. For the MCQ, both the experts' predictions and the actual correlations were available online in a published journal article and supplemental document, though behind the publisher's paywall (Furr et al., 2022). For the T-TCS, the paper by Reynolds et al. was not published; however, expert predictions and actual correlations were presented in a PsyArXiv technical report (https://osf.io/preprints/psyarxiv/brg9w_v1). For the P-React scale, no information was available online, aside from a preregistration that was not publicly viewable.

Second, the three validation studies differed from each other in terms of the conceptual depth provided in the instructions to the hypothesis-generation task. When making predictions for the MCQ, experts (and subsequently AI) received the names of the focal constructs (e.g., Global Morality) and the items on the MCQ's subscales, along with other information and instructions. For an ideal test of construct validity, the question to experts (or AI) should be "if we have a good scale assessing construct X, how should that scale correlate with each criterion measure?" To answer that question, experts (and AI) do not need the focal scale's items. In fact, access to those items might reduce the value of the process. A better approach is to provide experts (and AI) with clear definitions, explanations, and/or descriptions of the focal construct rather than scale items and/or simple labels for those constructs.[5] Such information can guide experts' and AI's understanding of the construct, may provide clearer insight into its likely nomological network, and may minimize misinterpretations potentially arising based on a simple scale label. Thus, the information provided to experts in the original validation work (and thus in our evaluation of AI) for the T-TCS and P-React withheld items on the focal scales and provided detailed conceptual definitions and descriptions of the focal constructs.

Actual Validity Correlations and Experts' Hypothesized Validity Correlations

The actual correlations between each scale/subscale and each criterion were computed from the previously collected

validation data. For example, Table 1's "Actual r" column presents the actual correlations for the Moral Character Questionnaire's Global Morality (MCQ-GM). Correlations for other scales/subscales are available in Furr et al.'s (2022) supplemental document (for other MCQ subscales), Reynolds et al. (2025), and Furr et al. (in preparation).

In original validations of each scale/subscale, experts produced hypotheses regarding validity correlations. As Furr et al. (2022) reported for the MCQ, five experts (faculty, postdoctoral fellow, and graduate students) independently considered each MCQ subscale's name and item content, reviewed item content from each of 21 criterion measures, and generated hypotheses regarding the correlation between the subscale and each criterion. Each expert thus independently provided seven sets of hypotheses, with 21 predictions per set. For example, Table 1's "Expert Predictions" columns present each expert's prediction regarding Global Morality and each criterion. Similar procedures were used to obtain experts' hypotheses for the T-TCS, with six experts considering 32 criteria, and for the P-React scale, with five experts considering 22 criteria. All expert hypotheses were generated prior to obtaining actual correlations.

Obtaining ChatGPT's Hypotheses

For each scale/subscale, we queried ChatGPT several times (i.e., trials) to match the number of original experts who provided hypotheses. For example, Table 1's "ChatGPT Hypotheses" columns present the MCQ-GM validity hypotheses produced by ChatGPT for each of five trials. In its first trial, ChatGPT predicted a correlation of $r = .80$ with MIQ:Self, $r = .75$ with MIQ:Integrity, and so on. Similarly, we queried ChatGPT six times for the T-TCS and five times for the P-React scale, reflecting the number of original experts for each.

Consistency of ChatGPT Hypotheses

For each scale/subscale, ChatGPT's hypotheses were highly consistent across trials. This can be seen at both the level of each criterion (across trials) and at the level of each trial (across criteria). At the level of individual criteria, ChatGPT's hypotheses varied little across trials. For instance, as Table 1 shows, hypothesized correlations between MCQ-GM and MIQ:Self ranged from .75 to .80, and hypotheses for each other's criteria ranged $\leq .20$ across trials for MCQ-GM.

At the level of trials, correlations between trials were extremely high (e.g., the correlation between Table 1's "Trial 1 and "Trial 2" columns). Viewing each trial as a rater, the average "inter-rater" consistency correlation is .99 for MCQ-GM (see Table 2's "Inter-rater Consistency: ChatGPT" column). This indicates that the criteria hypothesized to have strong positive (or strong negative, or near-zero) correlations with MCQ-GM in one trial were essentially exactly those hypothesized to have strong positive (or negative or near-zero) correlations in other trials. As Table 2 shows, similar consistency emerged for the remaining MCQ subscales, the T-TCS, and P-React. Average consistency correlations ranged between .95 and .99.

ChatGPT's consistency was generally greater than the consistency across experts. For example, as Table 1 shows, experts' hypotheses for the correlation between MCQ-GM and MIQ:Self ranged from .35 to .90 (recall, ChatGPT's hypotheses ranged from .75 to .80). The average inter-rater consistency correlation for experts was strong at .93 for MCQ-GM, and it ranged from .40 to .91 for the other scales/subscales (see Table 2).

We next aggregated hypotheses across ChatGPT's trials for each focal scale. For example, Table 1's "Aggregated Hypotheses: ChatGPT" column and Figure 2A display ChatGPT's average (across the five trials) hypothesized MCQ-GM correlation with each criterion. We view these aggregated hypotheses as ChatGPT's overall hypotheses regarding the correlations that should be obtained if MCQ-GM's scores are valid reflections of Global Morality. We conducted similar cross-trial aggregations of ChatGPT hypotheses and of experts' hypotheses for each focal scale (i.e., MCQ-Honesty, T-TCS, P-React, etc.). For example, Table 1's "Aggregated Hypotheses: Experts" column presents aggregations across the five experts' hypotheses for MCQ-GM's associations with each criterion.

Agreement Between ChatGPT's Hypotheses and Experts' Hypotheses

We next compared ChatGPT's aggregated predictions to experts' aggregated predictions, finding generally strong agreement, with some differences. Table 1's "Aggregated Hypotheses" columns and Figure 2B display ChatGPT's hypotheses alongside experts' aggregated hypotheses. As this table and figure show, ChatGPT's MCQ-GM hypotheses are largely similar to experts' hypotheses (e.g., MIQ:Self, $r = .77$, and $r = .69$, respectively).

In general, ChatGPT's hypotheses were somewhat "more positive" (i.e., farther above 0) than experts' hypotheses. For example, as Table 2 shows for MCQ-GM, ChatGPT hypothesized a correlation of .39 on average across criterion scales, while experts hypothesized a slightly weaker correlation of .33 on average. This arises from two facts: (a) ChatGPT's hypotheses were generally more extreme (i.e., larger in absolute magnitude) than experts' hypotheses, and (b) most criterion scales were predicted to be positively correlated with MCQ-GM. This pattern was observed for the other eight focal scales, as shown in Table 2.

ChatGPT's hypotheses were also slightly more varied across criterion scales, as compared to experts' hypotheses.

Table 1. Validity Correlations for the Moral Character Questionnaire's Global Morality Scale: Predictions From Artificial Intelligence, Predictions from Experts, and Actual Correlations.

	ChatGPT hypotheses					Expert hypotheses					Aggregated hypotheses			
Criterion scale	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5	Chat	Experts	Actual r	Sig difference In accuracy?
MIQ: Self	.80	.75	.80	.75	.75	.70	.70	.80	.35	.90	.77	.69	.71	Not Diff
MIQ: Integrity	.75	.80	.75	.70	.70	.90	.65	.80	.35	.90	.74	.72	.51	Not Diff
IPIP: Morality (TCI)	.70	.65	.70	.70	.80	.85	.55	.60	.25	.80	.71	.61	.65	Not Diff
MACS: Loyalty	.50	.30	.50	.50	.50	.40	.50	.60	.20	.70	.46	.48	.56	Not Diff
Honesty Questionnaire	.70	.60	.70	.65	.65	.70	.60	.60	.25	.60	.66	.55	.51	Exp Better
IPIP: Sincerity	.65	.55	.65	.60	.60	.90	.45	.40	.30	.30	.61	.47	.46	Exp Better
MACS: Honesty	.70	.55	.60	.65	.65	.80	.50	.50	.35	.40	.63	.51	.63	AI Better
IPIP: Fairness	.65	.50	.65	.55	.60	.90	.60	.60	.30	.50	.59	.58	.63	Not Diff
EPS: Egocentricity	−.60	−.70	−.60	−.50	−.60	−.70	−.60	−.70	−.35	−.60	−.60	−.59	−.43	Not Diff
EPS: Callousness	−.50	−.65	−.50	−.65	−.55	−.70	−.65	−.70	−.35	−.70	−.57	−.62	−.57	Not Diff
MACS: Fairness	.65	.55	.60	.60	.65	.60	.50	.40	.25	.50	.61	.45	.45	Exp Better
BFAS: Compassion	.55	.50	.60	.60	.60	.65	.60	.60	.30	.70	.57	.57	.64	Not Diff
IPIP: Empathy (TCI)	.65	.55	.55	.60	.55	.60	.65	.60	.25	.70	.58	.56	.64	Not Diff
MACS: Compassion	.65	.50	.50	.65	.50	.60	.50	.60	.25	.70	.56	.53	.57	Not Diff
Hedonic Behavior	−.50	−.40	−.30	−.40	−.30	−.20	−.20	−.40	.10	−.50	−.38	−.24	−.13	Exp Better
Profanity Use Freq	−.60	−.45	−.50	−.35	−.40	−.10	−.20	−.20	.00	−.20	−.46	−.14	−.11	Exp Better
Filial Piety	.55	.40	.60	.45	.55	.40	.30	.50	.25	.50	.51	.39	.41	Not Diff
BFAS: Politeness	.60	.45	.55	.55	.50	.25	.30	.30	.20	.40	.53	.29	.53	AI Better
Auth. Child-rearing Values	.50	.50	.40	.45	.45	.00	−.10	.00	.15	.20	.46	.05	.10	Exp Better
GACS: Fairness	.60	.50	.65	.55	.60	.80	.60	.60	.20	.60	.58	.56	.53	Not Diff
GACS: Honesty	.70	.60	.65	.65	.65	.85	.60	.60	.15	.60	.65	.56	.74	AI Better

Note: MIQ = Moral Identity Questionnaire; IPIP = International Personality Item Pool; TCI = Temperament and Character Inventory; MACS = Moral Actions and Concerns Scale; EPS = Expanded Psychopathy Scale; BFAS = Big Five Aspects Scale; Auth = Authoritarian; GACS = Global Assessment of Character Strengths; AI = Artificial Intelligence; Exp = Expert.

Table 2. Evaluating the Pattern Approach to Validity.

	Inter-"rater" consistency		Means			Standard deviations			Agreement	Accuracy		
Scale	AI	Exp	AI	Exp	Actual	AI	Exp	Actual		AI	Expert [95%CI]	Difference
Study 1 (Moral Character Questionnaire)												
GM	.99	.93	.39	.33	.38	.45	.40	.38	.95	.94	.97 [.89, .99]	NS
Loyalty	.97	.80	.26	.24	.36	.31	.25	.39	.94	.86	.86 [.45, .95]	NS
Honesty	.99	.84	.38	.28	.37	.47	.32	.39	.91	.94	.92 [.78, .97]	NS
Fairness	.99	.89	.38	.28	.37	.44	.35	.40	.93	.93	.95 [.79, .99]	NS
Compassion	.97	.91	.34	.29	.35	.42	.38	.35	.96	.96	.97 [.88, .99]	NS
Purity	.97	.40	.24	.15	.21	.36	.28	.27	.96	.93	.90 [.67, .96]	NS
Respect	.98	.74	.36	.21	.35	.40	.25	.37	.94	.91	.87 [.51, .96]	NS
Study 2 (Truthful Communication Scale)												
Truthful Comm	.99	.86	.16	.10	.12	.57	.38	.39	.93	.95	.96 [.91, .98]	NS
Study 3 (Patience Scale)												
Patience	.95	.73	.28	.16	.18	.44	.29	.34	.95	.88	.96 [.90, .98]	Experts

Note: GM = Global Morality; AI = Artificial Intelligence; Exp = Expert.

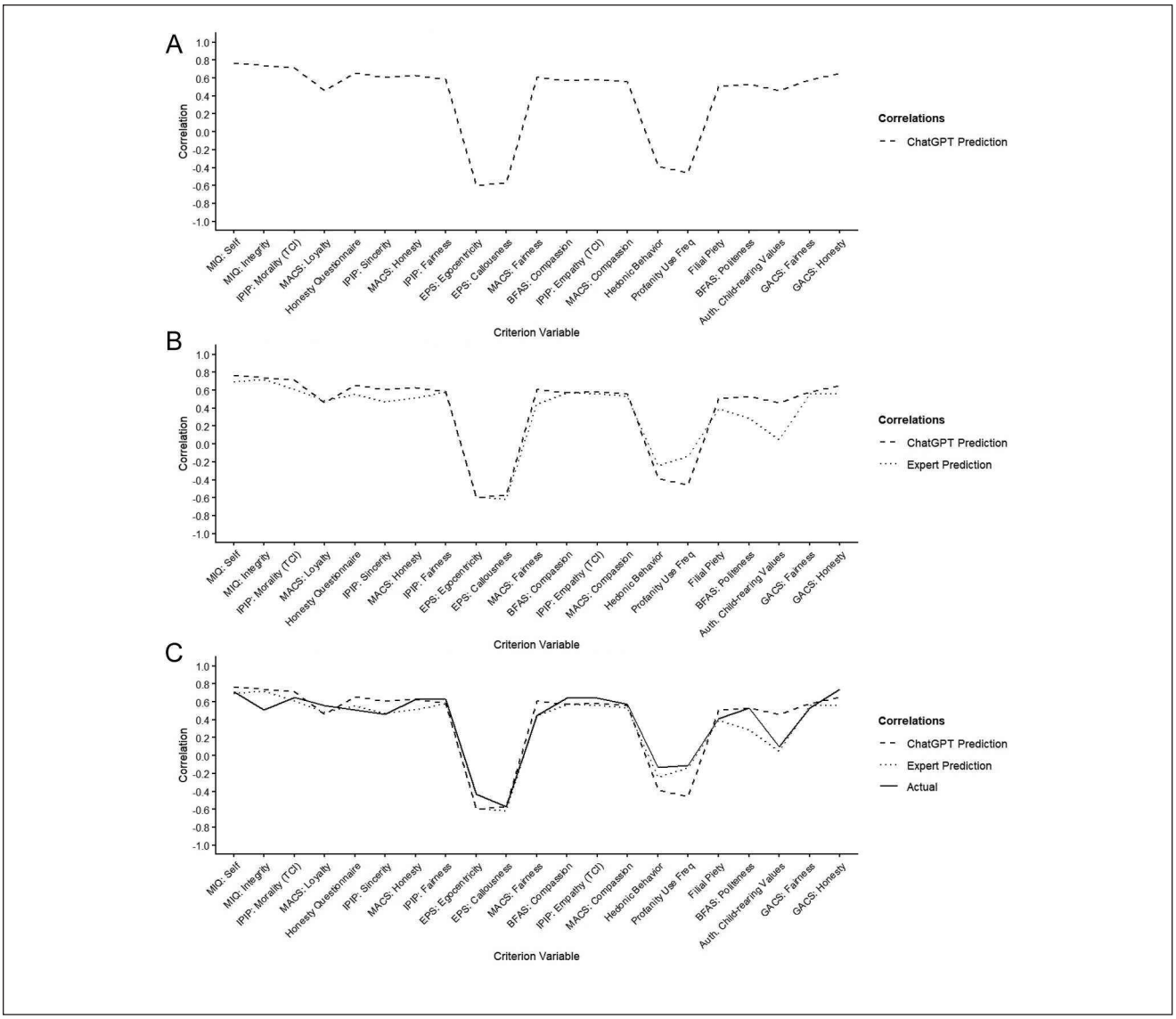

Figure 2. ChatGPT-Predicted, Expert-Predicted, and Actual Validity Correlations: Moral Character Questionnaire's Global Morality Subscale. (A) MCQ-GM ChatGPT-Predicted Correlations. (B) MCQ-GM ChatGPT-Predicted and Expert-Predicted Correlations. (C) MCQ-GM ChatGPT-Predicted, Expert-Predicted, and Actual Correlations.

For MCQ-GM as an example, Table 1 shows that ChatGPT's hypotheses across criteria ranged from −.60 to .77 (standard deviation = .45), while experts' hypotheses ranged from −.62 to .72 (standard deviation = .40). As shown by the standard deviations in Table 2, this difference was observed for the other eight focal scales as well. Thus, ChatGPT provided slightly more wide-ranging predictions than did experts.

Correlating ChatGPT's and experts' aggregated hypotheses across criteria, we found strong agreement. For example, the correlation between Table 1's two "Aggregated Hypotheses" columns was $r = .95$. As also illustrated by the fact that Figure 2B's lines parallel each other closely, this correlation indicates that ChatGPT and experts generally agreed upon which criterion scales would have the strongest positive, weakest, and strongest negative correlations with MCQ-GM (i.e., the relative ordering of the hypotheses, across criteria). As shown in Table 2's "Agreement" column, we found similarly strong agreement for the other eight focal scales/subscales, ranging from $r = .91$ to $r = .96$.

Accuracy of the Hypotheses

Finally, we examined the accuracy of ChatGPT's hypothesized validity correlations in terms of correspondence with

actual correlations. Consider Table 1's "Actual r" column, which displays the actual correlation between MCQ-GM and each criterion. How closely do ChatGPT's hypotheses correspond with these empirical values? We examined "correspondence with actual validity correlations" from two perspectives.

Pattern-Level Accuracy. Pattern-level accuracy is essentially the degree to which AI (and experts) correctly hypothesized the relative ordering of a set (or pattern) of correlations. High pattern-level accuracy means that a pattern of hypothesized correlations (i.e., which criteria were hypothesized to be relatively positive vs. weak vs. negative) corresponds well with the pattern of actual correlations. We quantify pattern-level accuracy as a correlation between a set of hypothesized correlations and a set of actual correlations, paralleling the "quantifying construct validity" (QCV; Furr & Heuckeroth, 2019; Westen & Rosenthal, 2003) procedure. This perspective is particularly useful when a large number of criterion measures are examined, and it has been used in scale validation (e.g., Costello et al., 2022; Miller et al., 2016; Moultrie & Engel, 2017; Porcerelli et al., 2016; Poythress et al., 2010; Suzuki et al., 2017; Wehner et al., 2021).

When accuracy is operationalized from this perspective, ChatGPT generally performs well. For example, for MCQ-GM, the correlation between ChatGPT's (aggregated) predicted correlations and the actual correlations was $r = .94$, as shown in Table 2. This result is again illustrated by the fact that the shape of ChatGPT's line in Figure 2C closely parallels the shape of the "Actual" correlation line, indicating that the criteria *hypothesized* by ChatGPT to have relatively strong positive (or weak or negative) correlations were almost exactly those that *actually* had relatively strong positive (or weak or negative) correlations with MCQ-GM. For comparison, we also correlated experts' (aggregated) predicted correlations with the actual correlations, with $r = .97$ (see Table 2 and Figure 2C).

Traditional parametric procedures are not well-suited for testing the significance of the difference between ChatGPT and expert accuracy correlations, so we used a bootstrapping procedure via the *boot* package in *R* (v1.3-31; Canty & Ripley, 2024).[6] We generated a bootstrapped 95% confidence interval around the experts' accuracy correlation (see Table 2) and considered ChatGPT's accuracy significantly different if it fell outside that interval. This revealed that the difference in pattern-level accuracy for MCQ-GM predictions was not statistically significant.

As Table 2 shows, results were similar for the other eight focal scales/subscales. ChatGPT's accuracy ranged from .86 to .96, experts' accuracies ranged from .86 to .97, and the difference was nonsignificant for all but one scale. For the P-React scale, ChatGPT's accuracy was high ($r = .88$), but was significantly lower than experts' accuracy ($r = .96$).

Criterion-Level Accuracy. A second way to examine accuracy is separately for each criterion. For a given criterion scale, accuracy is the degree to which its hypothesized correlation matches its relevant actual correlation. For example, ChatGPT hypothesized that the MCQ-GM should correlate $r = .77$ with MIQ:Self. Does that match the actual correlation between MCQ-GM and MIQ:Self? This perspective is likely most relevant when one or a few criterion measures are examined, and it reflects the absolute match between values (rather than reflecting similarity in relative magnitudes across correlations, as in the pattern-level perspective). When conducted for each criterion scale, we obtain a criterion-level perspective on AI's (and experts') accuracy.

Consider again the aggregated hypotheses in Table 1, for MCQ-GM. ChatGPT hypothesized the correlation between MCQ-GM and MIQ:Self to be $r = .77$, experts hypothesized the correlation to be $r = .69$, and the actual correlation was $r = .71$. Thus, ChatGPT's hypothesis was somewhat further from the actual correlation (indicating greater inaccuracy) than was experts' hypothesis (inaccuracies of .06 and −.02, respectively) as compared to experts' hypothesis. This pattern was observed for MCQ-GM, on average, across criteria. As Table 3 shows, the average absolute value of inaccuracies was somewhat larger for ChatGPT than for experts (.12 vs. .08, respectively).

Again unfortunately, traditional parametric tests are not suited to evaluating the statistical significance of such differences.[7] Therefore, we implemented a four-step procedure to determine whether ChatGPT's hypothesis for a given criterion was significantly more or less accurate than the experts' hypothesis for that criterion. First, we calculated the relevant standard error of each actual correlation (e.g., for MCQ scales, $N = 246$, so standard error = .064). Second, we Fisher-transformed each set of three correlations (i.e., the actual correlation, ChatGPT's hypothesized correlation, and experts' hypothesized correlation) and computed the raw discrepancy between the (transformed) actual correlation and each (transformed) hypothesis (e.g., raw discrepancy = .133 and −.039 for ChatGPT and experts' predictions for MCQ-GM and MIQ:Self, respectively). Third, we divided raw discrepancies by the relevant standard error, obtaining an error-calibrated discrepancy for each hypothesis (i.e., 2.08 and .610, for ChatGPT and experts). Fourth, based on the logic of declaring statistical significance when a statistic is (approximately) two standard errors away from a null value, we considered ChatGPT's hypothesis significantly more accurate if (the absolute value of) its error-calibrated discrepancy was itself two or more standard errors lower than the (absolute value of) experts' error-calibrated discrepancy. Similarly, we considered the experts' hypothesis significantly more accurate if its error-calibrated discrepancy was two or more standard errors lower than ChatGPT's error-calibrated discrepancy, and we declared the two not significantly different if the

Table 3. Evaluating the Accuracy of a Criterion-Specific Approach to Validity.

	Average absolute inaccuracy		Percent of criterion scales for which AI or experts were significantly more accurate		
Scale	AI	Experts	AI more accurate	Experts more accurate	No significant difference
Study 1: Moral Character Questionnaire					
Global Morality	.12	.08	14%	29%	57%
Loyalty	.21	.23	14%	10%	76%
Honesty	.13	.15	48%	38%	14%
Fairness	.12	.13	29%	29%	43%
Compassion	.09	.10	14%	24%	62%
Purity	.13	.11	19%	24%	57%
Respect	.11	.21	71%	24%	5%
Study 2: Truthful Communication Scale					
Truthful Comm.	.19	.09	9%	66%	25%
Study 3: Patience Scale					
Patience	.18	.08	9%	45%	45%

two error-calibrated discrepancies were within two standard errors of each other. For the MIQ:Self scale, ChatGPT's discrepancy was not significantly different from the experts' discrepancy (i.e., $2.06 - .610 < 2.00$). That is, ChatGPT's hypothesis was not significantly further away from the actual correlation than was the experts' hypothesis.

Using this procedure, we conducted a significance test for the MCQ-GM's association with each of its criterion measures. As Table 1's final column shows, ChatGPT's hypotheses for MCQ-GM were significantly better than experts' hypotheses for 3 of the 21 criterion scales (14%), experts' hypotheses were significantly better than ChatGPT's hypotheses for 6 scales (29%), and neither was significantly better or worse than the other for 12 scales (57%).

We applied this procedure to all nine focal scales/subscales. As Table 3 summarizes, neither ChatGPT nor experts consistently outperformed the other. ChatGPT's hypotheses were more frequently accurate than experts' hypotheses for some scales; they were more frequently less accurate for other scales, and many comparisons revealed nonsignificant differences in accuracy. This was particularly the case among MCQ subscales, which also varied in terms of which source (ChatGPT or experts) tended to have better accuracy. ChatGPT seemed to perform somewhat worse than experts for the T-TCS and P-React scale, though many of those comparisons were nonsignificantly different. Specifically, ChatGPT was significantly more accurate than experts for only 9% of the criteria for T-TCS and P-React. In contrast, it was significantly *less* accurate for 66% (for T-TCS) and 45% (for P-React) of the criteria.

Replication With Gemini

To evaluate whether these results are unique to ChatGPT 4o, we replicated all analyses using a different AI chatbot, Google Gemini. Available in Tables S1 and S2 in the Online Supplemental Material, results indicate that Gemini performed similarly to ChatGPT. Gemini's inter-trial consistency ranged from .85 to .99 (M = .95), which was slightly lower than ChatGPT's (M = .98) but somewhat higher than experts' (M = .79). On average, its hypotheses (mean hypothesized correlation = .27 across all focal scales and criteria) were slightly less strongly positive than ChatGPT's (M = .31) but slightly higher than experts' (M = .23). Its agreement with experts' hypotheses ranged from .87 to .98 (M = .94), which was similar to ChatGPT's agreement with experts (M = .94).

Gemini's accuracy was very similar to ChatGPT's. From a pattern-oriented perspective, Gemini's accuracy ranged from .79 to .97 (M = .93, ChatGPT M = .92). For eight of the nine focal scales, its pattern-oriented accuracy was not significantly different from experts' accuracy. As with ChatGPT, the lone exception was P-React, for which Gemini's accuracy was significantly lower than experts' accuracy (.79 vs .96). From a criterion-specific level, Gemini's accuracy was more similar to experts' accuracy than was ChatGPT's accuracy. Across all focal scales and all criteria, Gemini's hypothesized correlations were significantly closer (than experts' predictions) to actual correlations 20% of the time, they were significantly further (than experts' predictions) from actual correlations 21% of the time, and they were not significantly closer or further 59% of the time. The corresponding values for ChatGPT were 25%, 32%, and 43%.

Discussion

There is great potential value to a scale validation process guided by precise a priori hypotheses regarding validity correlations, but researchers rarely adopt this approach. AI might help researchers navigate two challenges that may be limiting the use of such validation procedures. We

evaluated whether AI can help researchers to identify relevant criterion measures and to generate precise a priori hypotheses for testing convergent and discriminant validity. Findings suggest that AI can indeed facilitate scale validation by addressing these challenges.

AI Can Help Identify Relevant Criterion Measures

To evaluate a scale's validity across criterion measures, researchers must identify and select a range of theoretically relevant measures. This can be challenging, as some constructs might not initially appear to have many clear theoretically relevant potential correlates, or as researchers might simply be unaware of relevant constructs and measures. We found that ChatGPT 4o can competently assist with this process, as it provided suggestions that converged well with criterion measures that were generated independently by researchers. It did so in a fraction of the time initially required by researchers.

Thus, AI can serve as a tool to generate ideas for identifying constructs and measures to use in evaluating criterion, predictive, convergent, or discriminant validity. Based on a script (see our example) that describes a focal construct and requests ideas, AI can facilitate such identifications rapidly. Even if it does not fully replace researchers' own ideas regarding relevant constructs and measures, AI can quickly supplement or expand those ideas. It might well produce options that failed to appear among researchers' ideas, potentially improving the scope and quality of validation work with minimal costs in time or effort. That said, researchers should expect some of AI's suggestions to be problematic in some way (e.g., incorrectly titled or accompanied by an incorrect citation), and a small number simply might not exist.

AI Can Help Generate Precise Hypotheses for Convergent and Discriminant Validity

A highly rigorous hypothesis-driven approach to scale validation benefits from precise a priori hypotheses regarding validity correlations. Imprecise or post hoc hypotheses provide weak tests of a scale's validity, as they open a door for subjective, impressionistic, and potentially biased conclusions. However, producing and testing precise a priori predictions might seem risky or daunting for many researchers. A useful approach can be to rely on panels of experts to produce such hypotheses, but doing so imposes on those experts and may take significant time.

We found that AI can quickly produce precise and plausible hypotheses regarding convergent (or criterion or predictive) and discriminant validity. There were at least three key findings. First, the "inter-rater consistency" of hypotheses obtained from different trials using AI was extremely strong, even somewhat stronger than consistency across experts. AI avoided some idiosyncratic mistakes or misinterpretations that sometimes occur with conscientious experts (e.g., an expert occasionally specifies the magnitude of a hypothesized correlation but forgets to specify the direction). Second, AI's hypotheses generally agreed strongly with experts' hypotheses, at least in terms of the pattern of correlations across criteria. This suggests that AI's hypotheses are generally good approximations of experts' hypotheses, in terms of which criteria will be positively, weakly, and negatively correlated with a focal scale. Third, AI's hypotheses were roughly as accurate as experts' hypotheses, in terms of corresponding with actual validity correlations. Although accuracy varied somewhat by perspective (pattern-level or criterion-level) and by focal scale, most comparisons revealed no significant differences between AI's and experts' accuracy. Similar results emerged for both ChatGPT 4o and Google Gemini, both producing psychologically reasonable hypotheses in a fraction of the time required by experts.

We thus conclude that researchers can use AI (at least ChatGPT 4o and Google Gemini) to help generate precise a priori hypotheses for scale validation. We imagine several options or circumstances that researchers might consider. First, researchers might rely solely on AI to produce psychologically plausible hypotheses, completely bypassing expert input. Although we are not yet convinced that this is always advisable, it represents a practical trade-off. Researchers can weigh the risk of AI producing suboptimal hypotheses against the time and effort saved by relying on AI rather than on panels of experts. Few, if any, of our comparisons indicate that AI's hypotheses were dramatically worse than experts' hypotheses, and in fact, AI's hypotheses were not infrequently more accurate than experts' hypotheses. When combined with the time and inconvenience (to one's colleagues and acquaintances) saved by using AI, this might lead researchers to conclude that the risk of suboptimal hypotheses is small enough (or the likelihood of reasonable hypotheses is large enough) that relying fully on AI is appropriate. Similarly, some researchers might lack access to panels of experts to impose upon, and they thus might use AI to approximate experts that are otherwise unavailable. In our view, these are justifiable decisions. However, researchers might alternatively conclude that the time and inconvenience saved by relying fully on AI is not worth the risk (as low as it might be) of suboptimal hypotheses. Again, this can be a justifiable decision, depending on the researcher's circumstances.

A second option would be to use AI for hypotheses to consider alongside experts' hypotheses, rather than fully replacing experts. Individual researchers who lack access to (or do not wish to burden) panels of experts might generate their own hypotheses and use AI to produce complementary hypotheses. The researcher could use AI's hypotheses as a

second expert to be aggregated with their own hypotheses. Alternatively, the researcher could use AI as a way to confirm or question their hypotheses. For example, the researcher might initially hypothesize that a good measure of patience should be correlated with self-control at $r = .40$ and with impulsivity at $r = -.10$, but find that AI hypothesizes the correlations to be $r = .36$ and $r = -.48$, respectively. AI's hypotheses might lead the researcher to feel confident in her hypothesis for self-control, but to revisit and perhaps reconsider her hypothesis for impulsivity. Even researchers who can recruit a panel of experts might consider AI hypotheses alongside experts' hypotheses, using AI as one of several "experts" to be queried and aggregated. Finally, when using AI for any of these purposes, researchers might opt to query more than one AI platform, considering that each may have its own model, its own training, and its own information base.

The current work used actual correlations as standards to evaluate the accuracy or quality of AI's (and experts') hypotheses; however, we should acknowledge that this reverses the perspective in actual scale validation work. In validity studies, theory-based hypotheses (either from experts or AI) reflect the correlations that should be obtained if a focal scale is a good measure of its intended construct. Those hypothesized correlations are the standard, and the scale's quality is evaluated in terms of how well its actual correlations correspond to those hypotheses. In contrast, our current analyses used actual correlations as the standard and evaluated the quality of the "theorists" (i.e., experts, ChatGPT, and Gemini) in terms of how well their hypothesized correlations correspond to those actual correlations. In both contexts, a lack of convergence between hypothesized and actual correlations could reflect either (or both) of two possibilities. The current study's accuracy is based on the assumption that it reflects theoretically incorrect hypotheses (e.g., experts or AI-generated, poorly informed hypotheses). In contrast, most actual scale validation work is based on the assumption that a lack of convergence reflects theoretically incorrect actual correlations (e.g., the focal measure is a poor indicator of the intended construct, so its correlations are poor reflections of the way that a good measure should correlate with other variables). Our interpretation of convergence as accuracy (and of lower convergence as less accuracy) should be understood in that context.

Implications for Understanding AI

Although not the focus of this paper, our findings may have implications for understanding how AI stores and uses knowledge. With a generally high level of accuracy, AI hypothesized validity correlations between dozens of specialized psychological constructs. One potential explanation for this finding might be that AI had previously seen and stored those correlations, and it simply produced them from "memory." This explanation is based on the view that AI mainly operates by recording and then reproducing recorded information. However, this explanation seems unlikely. First, for AI to have stored the correlations available from the two relevant papers, it would have to have stored every number in the paper, including *p*-values, page numbers, means, demographics, or it would have had to make judgments about which numbers were important prior to storage. Second, AI did not predict the actual correlations, so there would have to be an explanation for the noise in a table lookup model. But most importantly, the results for P-React were not available online (though again, experts' hypotheses were preregistered but not publicly available); however, it was able to predict those accurately as well.

Rather, these findings are more consistent with the view that AI processes the information it encodes and stores it as abstract concepts and the interrelationships between them (e.g., Pavlick, 2023; Xu et al., 2023). If AI did not have the correlations in its memory, then it must have somehow inferred the correlations between the concepts from the information it did have. And it had to have done so in a way that was accurate. Although this study was not designed to investigate this process, its findings suggest that this process has certain features. Specifically, AI must have had: (a) a fairly accurate representation of patience (and the other focal scales), (b) a fairly accurate representation of the criterion concepts, and (c) an ability to predict the two concepts' likely co-occurrence in humans. It had to use (c) to estimate the degree of psychological connection between (a) and (b), each of which must have been fairly accurately abstracted. These concepts may be stored as explicit definitions, as pointers/vectors to other, related concepts, or in some other form. Nonetheless, AI's predictive performance is difficult to explain without positing some form of abstract representation of the focal concepts and of the concept of psychological relationship. Rather than just a stochastic prediction machine, AI appears to be a stochastic prediction machine based on an abstracted encoding of human psychology.

Limitations and Future Directions

AI is evolving rapidly, with capabilities frequently expanding. Based on our findings and the fast pace of expansion in AI, its usefulness in supporting hypothesis-driven evaluations of convergent and discriminant validity is unlikely to diminish. As AI models continue to improve in reasoning and language understanding, their ability to serve as valuable tools in construct validation will likely remain stable or even strengthen.

The current work has limited generalizability in at least three ways. First, we evaluated both ChatGPT 4o and Google Gemini, and the results were consistent across those

models. Nevertheless, further investigation may evaluate additional models (Anthropic's Claude, Deepseek, etc.) to reveal whether any are particularly well-suited (or poorly suited) to these tasks. Second, although we replicated results across nine scales/subscales (conducting essentially nine examinations of both ChatGPT 4o and Gemini), this set is limited. All focal scales assess constructs in moral psychology, and the generalizability of our findings to other types of constructs and scales is unknown. Third, the actual correlations from all nine focal scales are based on participants from the United States. Although they varied in age, gender, and recruitment strategy (i.e., some samples were from college students and some from online recruitment platforms), they are geographically and culturally limited to the United States.

Aside from addressing those limitations to generalizability, future research can explore additional issues. We devoted considerable time to tailoring and testing our scripts (e.g., with and without criterion scale names), but scripts might be revised in new ways that improve AI's performance. In addition, we provided no training to the AI models; instead used basic models likely to be most familiar and rapidly usable for most researchers. That said, some form of training (e.g., providing manuscripts that detail relevant theoretical perspectives or scales) might improve AI's performance, though this would increase the time spent on using AI. Finally, many of the constructs that we examined (i.e., morality, compassion, truthfulness, etc.) may have considerable representation in empirical work and other sources available to AI. It is unclear whether AI would perform as well when asked to generate hypotheses for constructs that are not as well-represented in relevant sources. Indeed, AI's worst performance was observed for patience, which is likely less extensively represented in the empirical literature than many or most of the other constructs that we examined.

Conclusions

Our goal has been to evaluate whether AI can facilitate a rigorous hypothesis-driven approach to construct validation in a way that may expedite and improve the validation process and thereby advance psychological assessment. Our findings suggest that AI may indeed be able to assist researchers by helping to identify relevant criterion measures and by quickly producing psychologically plausible, precise hypotheses for validity correlations. Using AI in these ways might accelerate and/or improve the process of scale validation.

Acknowledgments

The authors wish to thank the following individuals for serving as experts in the original validation work described here: Ashley Hawkins-Parham, Eranda Jayawickreme, Dillon Luke, Christian Miller, Mike Prentice, Juliette Ratchford, Caleb Reynolds, Carlos Santos, Ryan Smout, Emily Stokes, and Christian Waugh. The original validation studies were partially supported through grants from the John Templeton Foundation, the Templeton Religion Trust, and the Templeton World Charity Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation, Templeton Religion Trust, or the Templeton World Charity Foundation.

Data Availability

Relevant data (i.e., hypothesized correlations and actual correlations) are available at https://osf.io/k9ptg/?view_only=1bd312e312f44e15830a03c4404a89b8.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Data collection for this work was supported by the John Templeton Foundation, the Templeton Religion Trust, and the Templeton World Charity Foundation.

Ethical Considerations

This work is based partly on data collected for previous studies. The Institutional Review Board at Wake Forest University approved all data collection.

Consent to Participate

All participants indicated consent by agreeing (within online surveys) to participate after reading a consent form (before participating in the study).

Consent for Publication

Not applicable.

ORCID iD

R. Michael Furr https://orcid.org/0000-0002-4476-1907

Supplemental Material

Supplemental material for this article is available at https://osf.io/k9ptg/?view_only=1bd312e312f44e15830a03c4404a89b8.

Notes

1. Alternative views also highlight the importance of theory in defining and evaluating validity (e.g., Borsboom et al., 2004) and thus implicitly highlight the importance of theoretically guided hypothesis testing.
2. In early phases of validating the T-TCS, Reynolds et al. (2025) referred to the relevant construct as "moral truth speaking." We adopted that label when working with AI, as shown in the Online Supplemental Material.

3. We initially considered evaluating the freely available ChatGPT 3.5. However, initial testing of its ability to simply use item content to produce reasonable predictions between scales proved unsuccessful. This testing involved no constructs ultimately used for the study. Rather, it involved items and constructs related to the Big Five personality traits. Although ChatGPT 3.5 performed poorly in preliminary tests, ChatGPT 4o performed well. We thus focused on ChatGPT 4o rather than 3.5 for our core analyses.
4. Here, we focus on one of three samples in Furr et al.'s (2022) analysis of the MCQ, one of two samples in Reynolds et al.'s (2025) analysis of the T-TCS, and one of two samples and one of two forms of P-React from Furr et al. (in preparation). For each scale, original validation results were highly similar across samples (and across forms, for P-React). We assumed such similarity would be observed here and thus examined one sample for each scale.
5. The idea is that hypotheses should emerge strongly from the conceptualization of the focal construct. It treats the key question as "in what ways should a good measure of the focal construct (as we have defined and described it) correlate with each of these other measures." This keeps experts' (and AI's) focus on the intended construct, rather than on a potentially flawed set of items to be used as a measure of that construct. We likely do not want to know how a potentially flawed set of items should correlate with a set of criterion measures. Rather, we want to know how a good measure of the intended construct should correlate with those criteria. An alternative approach could also ask experts or AI to also consider methodological issues (e.g., reliability of criterion measures and similarity of method variance between the focal measure and criterion measures; Furr & Heuckeroth, 2019).
6. Parametric tests are problematic for these correlations because the "unit of observation" is the criterion scale. Criterion scales were not randomly sampled, which violates one assumption of parametric tests. Moreover, the inferential population is not clear, as the most relevant "population" would be some kind of population of criterion scales. Finally, these criteria are not "independent observations" in any typical sense, again violating an assumption and raising ambiguity about appropriate degrees of freedom.
7. Existing parametric tests can evaluate whether an observed correlation differs from a single (null) hypothesis value. However, we know of no test to evaluate whether an observed correlation is closer to one hypothesized value than to another hypothesized value, which is the question here.

References

Abdurahman, S., Vu, H., Zou, W., Ungar, L., & Bhatia, S. (2024). A deep learning approach to personality assessment: Generalizing across items and expanding the reach of survey-based research. *Journal of Personality and Social Psychology*, *126*(2), 312–331. https://doi.org/10.1037/pspp0000480

AERA, APA, & NCME. (2014). *The standards for educational and psychological testing*. https://www.testingstandards.net/open-access-files.html

Aquino, K., & Reed, A. II. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, *83*(6), 1423–1440. https://doi.org/10.1037/0022-3514.83.6.1423

Banker, S., Chatterjee, P., Mishra, H., & Mishra, A. (2024). Machine-assisted social psychology hypothesis generation. *The American Psychologist*, *79*(6), 789–797. https://doi.org/10.1037/amp0001222

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Canty, A., & Ripley, B. D. (2024). *boot: Bootstrap R (S-Plus) functions*. R Package Version 1.3-31. https://doi.org/10.32614/CRAN.package.boot

Costello, T. H., Bowes, S. M., Stevens, S. T., Waldman, I. D., Tasimi, A., & Lilienfeld, S. O. (2022). Clarifying the structure and nature of left-wing authoritarianism. *Journal of Personality and Social Psychology*, *122*(1), 135–170. https://doi.org/10.1037/pspp0000341

Dumas, D., Greiff, S., & Wetzel, E. (2025). Ten guidelines for scoring psychological assessments using artificial intelligence. *European Journal of Psychological Assessment*, *41*(3), 169–173. https://doi.org/10.1027/1015-5759/a000904

Fan, J., Sun, T., Liu, J., Zhao, T., Zhang, B., Chen, Z., Glorioso, M., & Hack, E. (2023). How well can an AI chatbot infer personality? Examining psychometric properties of machine-inferred personality scores. *Journal of Applied Psychology*, *108*(8), 1277–1299. https://doi.org/10.1037/apl0001082

Furr, R. M., & Heuckeroth, S. A. (2019). The "Quantifying Construct Validity" procedure: Its role, value, interpretations, and computation. *Assessment*, *26*, 555–566. https://doi.org/10.1177/1073191118820638

Furr, R. M., Prentice, M., Hawkins Parham, A., & Jayawickreme, E. (2022). Development and validation of the Moral Character Questionnaire. *Journal of Research in Personality*, *98*, 104228. https://doi.org/10.1016/j.jrp.2022.104228

Furr, R. M., Waugh, C. E., Good, R. N., Ye, C., Li, J., Miller, C. B., Cole, J., & Porth, A. (In preparation). Development and initial psychometric evaluation of the Patient Reaction Scale and the Patience Regulation Scale: Theoretically grounded and complementary scales with alternate forms.

Götz, F. M., Maertens, R., Loomba, S., & van der Linden, S. (2024). Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. *Psychological Methods*, *29*(3), 494–518. https://doi.org/10.1037/met0000540

Haines, N., Kvam, P. D., Irving, L., Smith, C. T., Beauchaine, T. P., Pitt, M. A., Ahn, W.-Y., & Turner, B. M. (2025). A tutorial on using generative models to advance psychological science: Lessons from the reliability paradox. *Psychological Methods*. Advance online publication. https://dx.doi.org/10.1037/met0000674

Hernandez, I., & Nie, W. (2023). The AI-IP: Minimizing the guesswork of personality scale item development through artificial intelligence. *Personnel Psychology*, *76*(4), 1011–1035. https://doi.org/10.1111/peps.12543

Kjell, O. N. E., Kjell, K., & Schwartz, H. A. (2024). Beyond rating scales: With targeted evaluation, large language models are

poised for psychological assessment. *Psychiatry Research*, *333*, 115667. https://doi.org/10.1016/j.psychres.2023.115667

Krumm, S., Thiel, A. M., Reznik, N., Freudenstein, J.-P., Schäpers, P., & Mussel, P. (2024). Creating a psychological test in a few seconds: Can ChatGPT develop a psychometrically sound situational judgment test? *European Journal of Psychological Assessment*. Advance online publication. https://dx.doi.org/10.1027/1015-5759/a000878

Lee, P., Fyffe, S., Son, M., Jia, Z., & Yao, Z. (2023). A paradigm shift from "human writing" to "machine generation" in personality test development: An application of state-of-the-art natural language processing. *Journal of Business Psychology*, *38*, 163–190. https://doi.org/10.1007/s10869-022-09864-6

Markey, P. M., Campbell, H., & Goldman, S. (2025). A framework for the initial phases of personality test development using large language models and artificial personas. *Journal of Research in Personality*, *118*, 104647. https://doi.org/10.1016/j.jrp.2025.104647

Miller, J. D., Lynam, D. R., & Campbell, W. K. (2016). Measures of narcissism and their relations to DSM-5 pathological traits: A critical reappraisal. *Assessment*, *23*, 3–9. https://doi.org/10.1177/1073191114522909

Moultrie, J. K., & Engel, R. R. (2017). Empirical correlates for the Minnesota Multiphasic Personality Inventory-2-Restructured Form in a German inpatient sample. *Psychological Assessment*, *29*, 1273–1289. https://psycnet.apa.org/doi/10.1037/pas0000415

Pavlick, E. (2023). Symbols and grounding in large language models. Philosophical Transactions. *Series A, Mathematical, Physical, and Engineering Sciences*, *381*(2251), 20220041. https://doi.org/10.1098/rsta.2022.0041

Porcerelli, J. H., Cogan, R., Melchior, K. A., Jasinski, M. J., Richardson, L., Fowler, S., Morris, P., & Murdoch, W. (2016). Convergent validity of the early memory index in two primary care samples. *Journal of Personality Assessment*, *98*, 289–297. https://psycnet.apa.org/doi/10.1080/00223891.2015.1107573

Poythress, N. G., Lilienfeld, S. O., Skeem, J. L., Douglas, K. S., Edens, J. F., Epstein, M., & Patrick, C. J. (2010). Using the PCL-R to help estimate the validity of two self-report measures of psychopathy with offenders. *Assessment*, *17*, 206–219. https://doi.org/10.1177/1073191109351715

Reynolds, C. J., Jayawickreme, E., Wheat, R., Stokes, E., Santos, C., Fleeson, W., & Furr, R. M. (2025). Honesty as truthfulness: New trait and state measures of truthfulness to advance research and theorizing on when and why people are honest. *European Journal of Personality*. Advance online publication. https://doi.org/10.1177/08902070251338199

Singh, S., & Singh, W. (2024). AI-based personality prediction for human well-being from text data: A systematic review. *Multimedia Tools and Applications*, *83*, 46325–46368. https://doi.org/10.1007/s11042-023-17282-w

Suzuki, T., Griffin, S. A., & Samuel, D. B. (2017). Capturing the DSM-5 alternative personality disorder model traits five-factor model's nomological net. *Journal Personality*, *85*, 220–231. https://doi.org/10.1111/jopy.12235

Tong, S., Mao, K., Huang, Z., Zhao, Y., & Peng, K. (2024). Automating psychological hypothesis generation with AI: When large language models meet causal graph. *Humanities and Social Science Communications*, *11*, 896. https://doi.org/10.1057/s41599-024-03407-5

Wehner, C., Maaß, U., Leckelt, M., Back, M. D., & Ziegler, M. (2021). Validation of the Short Dark Triad in a German sample: Structure, nomological network, and an ultrashort version. *European Journal of Psychological Assessment*, *37*(5), 397–408. https://doi.org/10.1027/1015-5759/a000617

Westen, D., & Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology*, *84*, 608–618. https://psycnet.apa.org/doi/10.1037/0022-3514.84.3.608

Wetzel, E., & Killisch, J. (2025). Testing convergent and discriminant validity using a priori defined hypotheses [Editorial]. *European Journal of Psychological Assessment*, *41*(4), 253–256. https://doi.org/10.1027/1015-5759/a000918

Xu, Q., Peng, Y., Nastase, S. A., Chodorow, M., Wu, M., & Li, P. (2023). *Does conceptual representation require embodiment? Insights from large language models* (No. arXiv:2305.19103). arXiv. https://doi.org/10.48550/arXiv.2305.19103