International Journal of Assessment Tools in Education
2026, Vol. 13, No. 1, 224–247
https://doi.org/10.21449/ijate.1788563
journal homepage: https://dergipark.org.tr/en/pub/ijate
Research Article

Using large language models to evaluate ethical persuasion text: A measurement modeling approach

Matt Barney[1], **Stefanie A. Wind**[2*], **Vaishak Krishna**[3]

[1]TruMind.ai, USA
[2]University of Alabama, Department of Educational Studies, Tuscaloosa, AL, USA
[3]Apple, Inc., USA

ARTICLE HISTORY
Received: Oct. 7, 2025
Accepted: Dec. 7, 2025

KEYWORDS
Performance assessment,
Large language models,
Rater-mediated assessment,
Rasch measurement theory.

Abstract: As AI becomes prevalent in all stages of assessment procedures, it is essential to develop procedures to ensure that its use supports ethical and psychometrically defensible measurement. In this study, we consider how measurement principles can be directly incorporated into an ethical reasoning performance assessment in which Large Language Models (LLMs) serve as raters. We demonstrate how a measurement approach can be used to obtain defensible measures of LLM-generated text related to ethics, prompts designed to elicit text-based ethical persuasion responses, and individual learners. We demonstrate how measurement quality indicators can serve as guardrails to help mitigate potential AI-related risks that can impact learners, such as hallucinations or errors. We describe a novel approach to designing, implementing, and evaluating performance assessments with AI, with the goal of enabling effective personalized learning experiences

INTRODUCTION

The emergence of Large Language Models (LLMs) presents exciting opportunities and important challenges related to educational measurement. On one hand, these artificial intelligence (AI) systems have the potential to revolutionize the way we develop, administer, and interpret assessments, enabling more efficient, accurate, and personalized evaluation of student learning (Chia *et al*., 2023). The use of LLMs in educational measurement also raises concerns related to explainability, as well as the presence of hallucinations and bias in AI-generated content (Schwartz *et al*., 2022; Solaiman *et al*., 2023). These issues pose significant challenges and opportunities for future research to ensure validity, reliability, and fairness of AI-assisted assessments.

An active area of research is the use of LLMs as automated raters to evaluate examinee performances (e.g., Mizumoto & Eguchi, 2023). Although automated scoring is not new (Page, 1966), the recent advancements and availability of LLMs have inspired increased interest and

*Corresponding author at: Stefanie A. WIND swind@ua.edu Box 870321 Tuscaloosa, AL 35487, USA

e-ISSN: 2148-7456

applications of automated scoring in both research and practice, particularly within the context of writing assessment (e.g., Pfau *et al*., 2023; Ramesh & Sanampudi, 2021; Seßler *et al*., 2025). As in traditional rater-mediated assessments with human raters (Engelhard & Wind, 2018), LLM output is affected by a range of variables (e.g., prompt quality) that must be understood and managed to ensure the validity, reliability, and fairness of the results. Researchers have evaluated the quality of LLM ratings, often focusing on comparability with human raters (e.g., Floden, 2025; Kim *et al*., 2024; Wetzler *et al*., 2024; Yavus *et al*., 2024; Yun, 2023) as a source of validity evidence. Studies are less common that have used measurement modeling techniques to evaluate LLM ratings (e.g., Chan *et al*., 2023; Yamashita, 2024). As we discuss later in the manuscript, analyses informed by a measurement framework can provide researchers with clear guidelines for evaluating AI-assisted assessment processes beyond indicators of reliability and agreement. To the best of our knowledge, this application of measurement techniques to evaluate AI-assisted performance assessments has not yet been explored in the context of ethical persuasion assessment. In the current study, we use a performance assessment in the context of ethical persuasion to explore how measurement practices and principles can be used to evaluate and facilitate AI-based performance assessments in which LLMs act as raters.

Purpose

The purpose of this study is to demonstrate an approach to using measurement theory to evaluate LLM-generated ratings in a performance assessment in the context of ethical persuasion. LLM raters scored a sample of responses with specifically generated or selected characteristics, and we considered the quality of those ratings from a psychometric perspective.

With an emphasis on invariant measurement and the central role of substantive theory, we considered four research questions. Our first and overarching research question focused on the use of measurement models to evaluate LLM ratings:

1. How can measurement models be used to evaluate the psychometric quality of LLM-generated ratings in an ethical persuasion performance assessment?

We also considered the following additional questions to structure our analyses and discussion:

2. To what extent do LLM ratings exhibit useful psychometric properties related to model-data fit and rating scale functioning?
3. To what extent do various settings and prompt specifications impact the quality of LLM ratings?
4. To what extent do LLM ratings exhibit bias, as reflected in non-invariant calibrations between subgroups?

Our study contributes to previous research in several ways. First, broadly speaking, as LLMs play an increasingly important role in educational assessment, it is crucial to understand the variables that influence their behavior in a variety of contexts. Our study offers insight into variables that influence LLM ratings in an ethical persuasion assessment settings. Second, our study contributes to the emerging and rapidly developing body of research on LLM-based scoring for performance assessments. Our work specifically builds on the relatively limited set of research using measurement techniques to evaluate LLM ratings (e.g., Chan *et al*., 2023; Yamashita, 2024) and offers an exploration in a distinct disciplinary context. Third, our study contributes to the literature on performance assessment in ethical persuasion—demonstrating an AI-assisted approach to facilitating evaluations in this context.

Theoretical Framework for Ethical Persuasion Assessment

Our study used a performance assessment related to ethical persuasion as a context in which to explore measurement modeling approaches for evaluating LLM ratings. In this section, we describe the theoretical framework that guided the development and interpretation of scores from the performance assessment.

Our ethical persuasion assessment was designed to reflect Cialdini's (2009, 2014, 2016) framework of ethical persuasion and influence. A central component of this framework is that ethical persuasion requires educating a target of influence using information that is truthful, naturally fitting to the situation, and mutually beneficial (Barney, 2019). Cialdini's framework includes seven principles of persuasion. First, *reciprocity* suggests that people are more likely to say "yes" to requests from those who have previously provided them with a favor or gift. Second, *liking* refers to the tendency of individuals to be more easily persuaded by people they find attractive, similar to themselves, or cooperative. Third, *unity* describes the increased influence that comes from a shared identity between the persuader and the target. Fourth, *social proof* highlights the power of consensus, as people are more likely to engage in behaviors they see others similar to themselves performing. Fifth, *authority* refers to the tendency to follow the guidance of those perceived as trustworthy experts. Sixth, *consistency* describes the desire to remain consistent with one's values, promises and prior choices. Finally, *scarcity* suggests that opportunities are seen as more valuable when they are less available or limited in time, especially if competitors might take them away first. The *contrast phenomenon*, which underlies all these principles, emphasizes the importance of presenting information in a way that sharpens the differences between options (e.g. how much worse is the option I'm not recommending).

As shown in Table 1, we can conceptualize a progression of ethical persuasion from low to high that reflects varying degrees of adherence to the Cialdini principles. The progression in Table 1 reflects theories related to the development of ethical persuasion based on Commons Model of Hierarchical Complexity (MHC; Commons, 2007) and Fischer's Skill Theory (Fischer, 1980). Each principle has multiple amplifiers that could be present in a given situation, and experts exhaustively use all that are naturally present. At the lowest level, people have to honestly use at least one principle at a time, using methods to activate and amplify them. At the highest level, people systematically integrate Cialdini skills to transform their work and life.

Connecting Cialdini's framework to measurement theory, the progression of ethical behavior shown in Table 1 can be interpreted as a *construct map* that illustrates the development of ethical behavior along a linear continuum (Wilson, 2005, 2023). Construct maps provide a detailed, hierarchical representation of a given construct, along with the observable behaviors that provide evidence of these attributes. In the context of ethical persuasion, the construct map reflects an important developmental progression that has practical implications for targeting instruction. Specifically, Cialdini's model emphasizes that the lowest level of ethical persuasion is avoiding being a "smuggler of influence," which undermines relationships and erodes trust in the social proof, authority, commitment, or scarcity information provided. Once ethical behavior is established, it is developmentally inappropriate to revisit those easy tasks. Instead, instruction can be aligned with a learner's Zone of Proximal Development (Vygotsky, 1978), which reflects the level between which a learner can achieve independently and what they can achieve with support. By linking measurements to developmental frameworks, we can ensure that students receive appropriately challenging tasks that facilitate growth and promote long-term well-being. We revisit this construct map later in the manuscript as we consider the results from our performance assessment.

Table 1. *Construct map for ethical ınfluence.*

Level	Universal Definition	Persuasion Example
Highest Possible	Eminence represents the pinnacle of human achievement, a level of excellence so profound that it commands widespread recognition and respect. This extraordinary distinction, akin to winning a Nobel Prize, is so rare and conspicuous that only a select few individuals or teams ever attain it in human history.	This level signifies grandmastery over Cialdini's methods, applied so intentionally and effectively over decades that others seek their trusted advice for life's toughest challenges. They may be sought after by senior and thought leaders, including Presidents, Prime Ministers, CEOs, Technical and Scientific leaders, for guidance in wisely using persuasion principles to solve wickedly complex social issues. What distinguishes the grandmaster is how they've been applying the principles as a sleuth so brilliantly for decades, with such broad recognition that they become famously sought after.
Very High	This represents the apex of proficiency in a given field, surpassing nearly all contemporaries. The individual's flawless performance can catalyze significant commercial or social transformations, thereby making a substantial impact.	This individual is renowned in their profession for providing advice that others actively seek and value. They foster a culture and climate around them to cultivate relationships, reduce uncertainty, and motivate action ethically. They may develop tools, such as a braintrust, to promote the consistent and ethical application of these principles.
High	This denotes an exceptional professional standard. In performance or skill domains, the individual is a virtuoso, exhibiting near-perfect execution. They also serve as a mentor, fostering growth and performance in others.	The agent has earned a stellar reputation within their social network for their brilliant, wise and systematic application of Cialdini's Principles. They guide less proficient stakeholders in the correct use of these principles, providing support, guidance and correction when needed.
Medium -High	This represents an advanced level of proficiency, typically honed through years of education, training, or practice. The individual seldom errs and possesses the ability to guide others towards improved performance.	The individual's consistent application of the principles and contrast phenomenon has begun to establish them as a trusted guide others wish to follow. They act as a supportive "wingman/woman" to peers proficient in Cialdini's methods.
Medium	This refers to a superior intermediate level, where performance exceeds basic requirements. The individual typically operates with a high degree of accuracy, making virtually no mistakes.	The agent proactively actively and effectively employs all Cialdini Methods, utilizing job aids and systematically seeks out peer support to counteract potential oversights in complex situations due an acute awareness of their inattentional blindness.
Medium -Low	This represents a moderate level where performance is satisfactory and acceptable, though not exceptional. The individual can perform competently, albeit with occasional errors.	The agent actively and ethically applies all principles and contrasts in the supermajority of situations. However, they may occasionally overlook certain activators or amplifiers in some contexts; or neglect powerful contrasts at times.

Level	Universal Definition	Persuasion Example
Low	This denotes a low level of proficiency, often seen in those still acquiring expertise. The individual frequently commits basic errors, resulting in performance that falls slightly short of acceptable standards.	The agent correctly applies some principles but may occasionally overlook others, including ways to begin or strengthen the persuasion process. The agent might also lapse in proactively seeking out all principles, or identifying ways to sharpen contrasts at times.
Very Low	This refers to a very low level of proficiency, typically exhibited by a novice in the process of learning a skill. The individual's performance falls below acceptable standards.	The agent applies only a single principle of persuasion or a single contrast that is natural to the situation without looking for the others that are also present, bungling the opportunity.
Extremely Low	This signifies an extremely low level of proficiency, often seen in individuals who have just embarked on learning a new skill. Their performance is currently below acceptable standards.	The agent does not apply any principle of persuasion or any contrast phenomenon.
Lowest Possible (Absolute Zero)	*"Absolute zero" represents the lowest possible level of an attribute, potentially unprecedented in human history. In terms of developable skills, the individual or team may lack any exposure to learning opportunities, resulting in performance that is significantly below acceptable standards.*	

METHOD

Our procedures involved three main steps. Our first step was data collection; this involved gathering ethical reasoning texts to be scored in our assessment procedure. These texts were the object of measurement (i.e., examinees) in our assessment process, and our sample included both human-authored and AI-generated texts. Second, we implemented our AI-assisted performance assessment by using LLMs to rate the responses using our scoring rubric and rating scale. Third, we analyzed and evaluated the resulting ratings of those responses using a measurement modeling approach. We discuss each of these steps in detail in the following section.

Step One: Data Collection

Our data collection procedure involved collecting or generating text related to ethical persuasion that would subsequently be scored using LLM raters. We used a mix of human-authored and AI-authored ethical persuasion texts because we were interested in exploring how certain characteristics of the text may influence the quality of LLM rater judgments. This approach can be viewed as conceptually similar to a simulation study in which researchers generate examinee responses under known conditions to evaluate a measurement technique.

Sample

As shown in Table 2, the sample of texts in our assessment procedure included human-authored text, GPT-4 summaries of human text, and AI-generated text samples created with Tree of Thoughts prompts and GPT-4 (Yao *et al*., 2023; described further below). We included this combination of human and AI-generated texts to ensure that our sample reflected a wide range of ethical persuasion levels and characteristics known to impact ethical persuasion. We sought out real human examples to reflect extreme levels of low and high leadership ethics in human history. Second, because large numbers of human samples are hard to procure, we used GPT-4 to intentionally over-engineer a disproportionate number of extremely ethical and unethical samples. Finally, we proactively made sure the sample had enough information about stylistic sex differences in text, as we had small numbers of women in our human sample.

Table 2. *Human and synthetic sample overview.*

Sample Type	# Samples	Rationale
GPT-4 generated - Extremely Low ethics (rating of 1)	100	Oversample extremes, to proactively minimize areas of high uncertainty usually limiting Rasch Analyses
GPT-4 generated - Extremely high ethics (rating of 10)	100	
GPT-4 generated - Medium ethics (2-9)	222	To achieve equally small levels of uncertainty with ordinary samples that are easier to obtain, we generated a smaller sample of levels 2-9 of ethical proficiency.
Human authored - All levels	54	Real text samples (e.g. speeches, interviews) that reflect historically high and low levels of ethics in high profile leaders (e.g. Prime Ministers, historical figures, Presidents and CEOs)
GPT-4 summary of human speeches	11	In order to test whether synthetic summaries of real human speeches could also be estimated. English translations of other languages make this an especially challenging test (e.g. Deng Xiaping's speech translated from Chinese after Tiananmen Square massacre)
Gibberish/Word Salad	3	Wanted to test robustness to a wide range of use-cases, including robustness to nonsensical inputs
Total	503	

Prompts to generate responses

For the AI-generated responses, we used carefully designed prompting procedures to create ethical persuasion texts. In our assessment procedure, the prompts can be viewed as a type of "item" on which we observed examinee responses. We used a mix of human-authored prompts and prompts engineered via GPT-4 with carefully selected characteristics. The first author composed the human-generated prompts based on his substantive expertise in Cialdini methods. While writing the prompts, he scrutinized the literature from various schools of philosophy, cultural anthropology, theology, sociology, computational linguistics, social psychology, industrial-organizational psychology, political science and economics to write 110 total pilot prompts focused on ethics to be tested. The first and third author generated additional, engineered prompts via GPT-4 to a) create synthetic samples; and b) reflect different levels of ethics using a variety of prompt authoring methods from computer science. These included:

- Zero Shot learning: Enables LLMs to infer the correct output for tasks on which they have not been explicitly trained (Sun *et al*., 2021)
- Few Shot learning: Training LLMs on a very small amount of example data, asking them to generalize to novel examples (He *et al*., 2023)
- Chain of Thought prompts: Written to guide the LLM to articulate intermediate reasoning steps that improves performance on complex tasks (Wei *et al*., 2022)
- Bag of Words prompts: Written based on pre-LLM research showing statistical representations where text is a set of unordered words, disregarding grammar and word order, and often used with Classical Test Theory methods (Boyd *et al*., 2022; Zhang *et al*., 2010)

Table 3. *Example Cialdini ethics prompts.*

Prompt Type	Facet	Example Prompt	Inspiration
Few Shot, BARS, Bag of Words evidence; One Expert-based	True	You're an extremely careful, conscientious moral philosophy expert system at assessing whether a sample is good at applying all of Robert Cialdini's Principles of Persuasion in an entirely ethical, honest and truthful way. You need to be extremely careful and select only a single integer where 10 is absolute perfection, and 1 is absolutely no application of his principle, return only an integer and no text where: 1=the text does not use any (zero) person plural pronouns (e.g. we, our, us, lets) anywhere the sample; 2=the text uses only one or two first person plural pronouns (e.g. we, our, us, lets) throughout the sample; 3=the text uses at least three or more, but very few first person plural pronouns (e.g. we, our, us, lets) throughout the sample; 4=the text uses four or more (but not many more) first person plural pronouns (e.g. we, our, us, lets) throughout the sample; 5=the text uses at least five, but a moderately low number of first person plural pronouns (e.g. we, our, us, lets) throughout the sample; 6=the text uses at least six but a moderate number of first person plural pronouns (e.g. we, our, us, lets) throughout the sample; 7=the text uses a fairly large number of first person plural pronouns (e.g. we, our, us, lets), more than an average or medium amount, but less than a high amount throughout the sample; 8=the text uses a high number of first person plural pronouns (e.g. we, our, us, lets) throughout the sample; 9=the text uses a very high number of first person plural pronouns (e.g. we, our, us, lets) throughout the sample; 10=The text uses an extremely high number of first person plural pronouns (e.g. we, our, us, lets) saturated throughout nearly every part of the sample. Err on the side of extreme conscientious, careful conservatism and when in doubt return low ratings unless there is overwhelming evidence for a higher rating. Only return a single integer from 1-10 and no text.	Boyd *et al*. (2022)

Prompt Type	Facet	Example Prompt	Inspiration
Zero Shot, Multiple Expertise: Comp. Linguist, Psycholinguist	Naturally Present	Act as a computational linguist and evaluate the given sample on a scale from 1 to 10, based on its consistency with the cultural and historical context in which it is situated.To assess cultural and historical consistency, consider how well the sample aligns with the cultural and historical factors that give rise to its context. A well-composed sample should accurately reflect the cultural and historical context in which it is situated, including the values, beliefs, customs, and practices of that context.Please exercise extreme caution when evaluating the sample and only provide a high rating if there is overwhelming evidence that it accurately captures cultural and historical consistency. When in doubt, rate the sample lower. Remember, a higher rating should be given to a sample that demonstrates a high degree of accuracy in capturing cultural and historical consistency, while a lower rating should be given to a sample that lacks this accuracy.To provide your rating, please enter a single integer score between 1 and 10, without any accompanying text.	Geisler (2008)
Zero Shot, Inter-disciplinary Expertise-based, synonyms	Wise	Act as a world expert on interdisciplinary conceptions of wisdom, and rate the provided sample on a scale of 1 to 10 of whether it is good at applying all of Robert Cialdini's Principles of Persuasion with the highest possible levels of wisdom. You need to be extremely careful, conscientious, and prudent in deciding the most appropriate level where 10 is absolutely perfectly wise, and 1 is totally unwise where:Assess the given text excerpt in terms of how well it conveys wisdom, considering the array of findings from science that describe wise acts. Take into account attributes such as the ability to reconcile paradox, moral uprightness, selflessness, defiance of internal and external pressures, pursuit of balance, risk-taking, aspiration to enhance the human condition, advanced cognitive ability, deep insights linking cognition and motivation, profound thinking, emotional stability, compassion for others, adept problem-solving, connection to nature, humor, and elderly behavior emphasizing long-term love for humanity. Provide a rating from 1 (low wisdom) to 10 (high wisdom), reflecting the ultimate integration of mind and character for the greater good of the human species. Return only one integer from 1-10 and no text.	Staudinger, U.M., & Gluck, J. (2011)

In addition to these elements, because researchers have noted improved rater judgments with behaviorally-anchored rating scales (BARS) and examples (Schwab, Heneman & DeCotiis, 1975), we included BARS and examples as additional prompt characteristics. For the prompts that used BARS, we hypothesized that we could use the essence of Wilson's construct maps (Wilson, 2005, 2023) to approximate Cialdini's levels of ethics at each level of ethics proficiency ordinally on a ten-category rating scale, within each prompt as well as between prompts. Table 3 shows sample prompts for each facet of Cialdini's framework.

Step Two: Scoring

Scoring was conducted using LLMs as raters. We tested our approach to using LLM raters in a pilot study in which we used GPT3.5 Turbo to rate our initial pool of text samples. Results from the pilot study indicated acceptable psychometric properties that supported a full study (e.g., evidence of strong reliability, high levels of separation, and model-data fit statistics within established limits). Please see the Appendix for a detailed description of our pilot study methods and results. Scoring for our full study involved having each LLM use a 1-10 scale informed by Michael Lamport Common's Model of Hierarchical Complexity for each of Cialdini's dimensions, similar to how human raters would traditionally rate behavior in an assessment center. Additional details about the scoring process and rating scale are given in the Appendix.

Step Three: Data Analysis

The third step in our procedure involved evaluating the quality of LLM ratings using psychometric methods. Reflecting our emphasis on practical use and theoretically sound interpretation of measurement results, we used Rasch measurement theory (Rasch, 1960) as our guiding measurement framework. A central component of Rasch measurement theory is model-data fit analyses, where responses are evaluated against strict criteria, including tests for adequate unidimensionality, local item independence, and invariance. The guiding perspective that characterizes Rasch measurement applications is that these requirements are not expected to be fully realized in practice. Instead, the requirements serve as "guardrails" against which observed response patterns can be evaluated for the purpose of understanding and improving measurement procedures.

Drawing from the literature on measurement modeling for rater-mediated assessments (e.g., Linacre, 1989), we used extended versions of polytomous Rasch models that incorporate explanatory facets related to LLMs and prompts. These models can be broadly classified as Many-Facet Rasch Models, and they can also be discussed within the explanatory Item Response Theory framework (de Boeck & Wilson, 2004). The MFRM has been successfully applied in machine learning research (Barney, 2019; Kennedy *et al*., 2020; Sachdeva *et al*., 2022), as well as in rater-mediated performance assessments across a wide range of disciplines (Engelhard & Wind, 2018). In the context of our current study, MFRMs provide a rigorous and efficient approach to evaluating the variables that affect LLM output.

Using Facets (Linacre, 2020), we applied a Partial-Credit (PC) model (Masters, 2018) formulation of the Many-Facet Rasch Model (PC-MFRM; Linacre, 1989) to obtain location estimates and evaluate the measurement properties of the LLM ratings. We used the PC model formulation so that we could examine rating scale functioning in detail specific to individual prompts, as reflected in our research questions.

Our model included the following facets. First, we included a facet for text samples, which served as the examinee. Second, we included a facet for prompt, which served as a type of item. Third, we included a facet for the LLM, which served as the rater who scored each sample on each prompt. We also included sample-related explanatory facets to examine the effects of sample characteristics, including the overall ethics level, sex, human vs. GPT4 generated responses, the use of a prefix, header, Chain-of-Thought (COT), examples, bars, and emoji. Sample-related facets were oriented positively such that higher ratings corresponded to higher measures. A concise representation of the model can be stated as in Equation 1:

$$ln\left(\frac{P_{n(x_j=k)}}{P_{n(x_j=k-1)}}\right) = \lambda_r - \left(\theta_n - \textstyle\sum_{facets}\gamma\right) - \left(\beta_j - \textstyle\sum_{facets}\eta\right) - \tau_{jk} \tag{1}$$

In this equation, λr is the location of LLM-rater n, θn is the location of sample n, and βj is the location of prompt j on the latent variable. The model includes a linear combination of explanatory facets associated with samples ($\sum_{facets}\gamma$) and prompts ($\sum_{facets}\eta$). For samples, explanatory facets included the specified ethics level and sex classification (male or female). For prompts, the explanatory facets include indicators related to the type of prefix, the number of personas (one or two), the use of Chain of Thought, the use of examples, the use of a BAR, and the use of content-relevant emojis. Finally, the τ parameter reflects the threshold between categories in the ordinal rating scale specific to prompt j. To provide a frame of reference for interpreting location estimates, we set the average locations for all facets to zero logits except for the sample facet, which was the object of measurement.

From the model results, we focused on four major indicators. The first three: (1) model-data fit; (2) rating scale functioning; and (3) differential functioning provide support for interpreting (4) parameter location estimates, including visualizations that provide insight into the effect of prompt and sample-related characteristics with respect to the construct.

Model-data fit analysis

We evaluated the psychometric quality of the LLM ratings using model-data fit analyses based on Rasch model residuals. In practice, researchers who use Rasch models typically focus on two residual summary statistics: inlier-weighted (infit) and outlier-weighted (outfit) mean-square error (MSE) statistics. These residual fit statistics help researchers identify individual elements in an assessment system for which response patterns may threaten the interpretability of model estimates. Infit *MSE* statistics are weighted to avoid the undue influence of outliers (see Equation 2), whereas outfit *MSE* is not weighted (see Equation 3):

$$\text{Infit } MSE = \frac{\Sigma((r_i - E_i)^2 * W_i}{\Sigma\, W_i} \tag{2}$$

$$\text{Outfit } MSE = \frac{\Sigma((r_i - E_i)^2 / E_i}{n} \tag{3}$$

where

r_i is the observed response,

E_i is the expected response,

W_i is the information weight, and

n is the number of observations.

High values of infit and outfit *MSE* indicate that the response patterns may be noisy and unpredictable, while low values indicate that the responses may be redundant (i.e., overfitting). These statistics can help diagnose and improve the quality of a measurement system.

Rasch analysts also perform other quality tests to make sure the measurement requirements are realized, including explicit empirical tests of unidimensionality (e.g. Principal Components Analysis of standardized residual correlations), and independence (e.g. inter-item residual correlations), and special tests designed to see how deviant raw data are from Rasch ideals.

Rating scale analysis

We also evaluated our assessment procedure for evidence of effective rating scale functioning. Rating scale functioning reflects the degree to which estimates of rating scale thresholds from a measurement model (e.g., the PC-MFRM) support meaningful interpretation and comparison of ratings between elements of an assessment procedure, such as raters or prompts (Linacre, 2004; Wind, 2022). We examined two major indicators of rating scale functioning. First, we checked the threshold estimates (τ) for evidence that lower scale categories (i.e., lower levels of ethical persuasion) corresponded to lower locations on the latent variable than those associated with higher levels. Evidence of expected threshold ordering supports an ordered interpretation of scores with respect to the construct. Second, we examined the magnitude of the distance between adjacent thresholds for evidence that the scale categories provided distinct information about levels of ethical persuasion. Small distances between adjacent thresholds indicate that scale categories may be redundant, whereas large distances may indicate a need for additional levels of distinction.

Differential prompt functioning

Finally, we evaluated the prompts for evidence of differential prompt functioning (DPF; i.e., bias) related to the sex specification for the sample. We conducted the DPF analysis by adding an interaction parameter between the prompt facet and sample sex facet. We interpreted a significant interaction term as evidence of potential bias that warrants additional exploration using pairwise analyses between each prompt and sex group.

Parameter locations

Finally, after ensuring defensible psychometric properties, we examined location estimates and corresponding Standard Errors (*SE*) for the facets in our analysis. We considered the overall targeting between the distributions of each facet as well as locations for individual elements within each facet (e.g., individual LLMs and prompt characteristics). We examined the degree to which the element locations were distinct using statistics such as reliability of separation, separation statistics, and strata (Linacre, 1989). We used Wright Map displays to illustrate and facilitate the interpretation of the location estimates.

RESULTS

Preliminary Analysis

We examined the overall measurement quality of our ratings using the PC-MFRM in a preliminary analysis. This analysis revealed several idiosyncrasies that warranted adjustments prior to our final analysis. With the original data, location estimates for the PC-MFRM explained 25.42% of the variance in the ratings. This value supports the interpretation of measures on a unidimensional scale (Engelhard Jr & Wang, 2020; Reckase, 1979). The average element-level model-data fit statistics were close to 1.00 for each facet; this value is typically considered evidence of adequate fit to Rasch measurement theory models (Smith, 2004; Wu & Adams, 2013). Although these results provide initial support for the interpretation of measures, close inspection revealed several idiosyncrasies that needed to be addressed prior to using these results to facilitate further analyses

Model-data fit

Relative to the other facets, the standard deviation (*SD*) values of model-data fit statistics for samples were relatively high—suggesting that there was a subset of samples for whom rating patterns may not contribute meaningfully to the measurement procedure. Recognizing the continuous nature of *MSE* fit statistics and the lack-of-agreed-upon critical values for these statistics (Wu & Adams, 2013), we used the 90th percentile of the empirical distribution of infit *MSE* (> 1.51) and outfit *MSE* (> 1.55) to evaluate model-data fit for our samples (Walker & Wind, 2020).

After removing 52 samples with extreme misfit, the overall proportion of variance explained by PC-MFRM measures was higher (26.19%), and average model-data fit statistics for all facets remained close to 1.00. However, we still identified some samples with relatively high *MSE* fit statistics. At this stage, we used a more conservative criterion and identified samples with fit statistics that exceed the 95th percentile of the empirical distribution for potential removal. This procedure identified an additional 34 samples for removal from the analysis.

Next, we identified prompts with extreme misfit and removed them from the calibration sample. As we did for samples, we used the 95th percentile of the empirical distribution of infit *MSE* (> 1.41) and outfit *MSE* (> 1.69) to identify prompts with extreme misfit. These statistics identified 21 prompts as candidates for removal. We also examined the point-measure correlations for these prompts and found that they were near zero or slightly negative—providing further support for removal. In a third analysis, model estimates explained 30.46% of the variance, average fit statistics were close to 1.00 for all facets, and the *SD* of model-data fit statistics was reduced for all facets.

Preliminary rating scale analysis

Results from our preliminary rating scale analysis revealed several idiosyncrasies related to rating scale category ordering. First, there were several pairs of thresholds that were nearly always disordered: Thresholds 2 and 1, Thresholds 4 and 3, and Thresholds 6 and 5. Second, the highest three thresholds were disordered less often among prompts compared to the lower thresholds. The frequency of disordered threshold pairs for individual prompts ranged from 1 to 6 (M = 3.75, SD = 0.90). Together, these results suggest that rating scale functioning may be improved by combining several of the sets of adjacent categories in the lower end of the scale (Tsai *et al*., 2024).

We iteratively tested several re-coding schemes, including revised scales with three, four, five, and six categories, and examined the results for overall patterns of rating scale functioning. Although there were still some prompts with disordered thresholds present, we found that a four-category solution seemed to provide the optimal balance of category functioning, model fit, and statistical information. We recoded the original scale such that original ratings of 1 or 2 were coded as 1, original ratings between 3 and 5 were coded as 2, original ratings equal to 6 or 7 were coded as 3, and ratings equal to or higher than 8 were coded as 4. Because the scale functioning was generally comparable across prompts with this revised scale, we changed the analysis model from a PC model formulation to an RS formulation (Andrich, 2018), where we estimated one set of thresholds for all prompts. We estimated the RS-MFRM with the revised scale and sample and dropped any remaining prompts and samples with substantial evidence of misfit.

Differential prompt functioning

The omnibus chi-square test for the interaction between prompts and sample sex was statistically significant ($\chi^2(576) = 1575.7$, $p < .01$). This result suggests that there may be differences in prompt difficulty between sex groups for individual prompts. Considering the relatively large sample size, we focused on an effect size measure of the magnitude of DPF in logit-scale units. Specifically, we examined the contrast statistic, which is calculated as the difference in prompt difficulty between sex subgroups on the logit scale. Following Engelhard and Myford (2003), we used a value of .30 logits to identify substantial bias. We iteratively removed 4 prompts with contrast statistics that exceeded 0.30 logits until there was no longer evidence of meaningful DPF. Three of the prompts showing bias were very short, and did not include a BARS or examples. One dealt with a conception of wisdom that may not be well represented in the LLMs to rate effectively.

Results with Final Scale and Sample

With this final sample and rating scale identified, we further refined our analysis to focus only on the facets that were most central to the purpose of our study and proposed use of our measurement procedure: prompt indicators, sample indicators, LLM indicators, sex (for DPF) associated with samples, and the ethics level associated with samples. We revised our final model to only include these facets.

With the final model, sample, and rating scale, the RS-MFRM model estimates explained 38.41% of the variance, average fit statistics were close to the Rasch ideal of 1.00 for all facets (see Table 4). The average inter-item standardized residual correlation was equal to r = .07. This near-zero correlation suggests adequate adherence to the local item independence requirement for the measurement model.

Table 4. *Summary of facet results from analysis with final sample, final model, and collapsed rating scale.*

Facet	*N* Elements	Reliability of Separation	Separation	Strata	Model-Data Fit Infit MSE		Outfit MSE		Location		*SE*	
					M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Sample	415	> .99	24.09	32.45	1.01	0.21	1.03	0.22	-0.06	0.61	0.02	0.01
Prompt	288	> .99	25.86	34.81	1.00	0.25	1.02	0.27	0.00	0.48	0.02	0.00
LLM	3	> .99	290.70	387.90	1.00	0.14	1.02	0.16	0.00	0.59	0.00	0.00
Sex	2	> .99	36.96	49.62	1.01	0.01	1.03	0.01	0.00	0.06	0.00	0.00
Ethics Level	3	> .99	107.50	143.40	1.02	0.07	1.04	0.08	0.00	0.25	0.00	0.00

Wright map

Figure 1 is a Wright Map that shows the calibration of the facets on a common logit scale. This figure is an abbreviated version of the complete Wright Map that shows only the sample and LLM facets to promote a clearer visual interpretation. The histograms on the left of the plot show the distribution of elements of the samples and LLMs. Location estimates for individual samples ranged from θ = -1.79 for the sample with the lowest ratings to θ = 1.83 logits for the sample with the highest ratings. The average sample location was equal to M_θ = -0.06 logits. To further explore the calibration of our samples, we examined the location estimates for a subset of individual text samples from senior leaders, presidents, and prime ministers who we expected to vary with respect to ethical persuasion levels according to our theoretical framework. As shown in Figure 2, location estimates for these samples aligned with our a priori expectations.

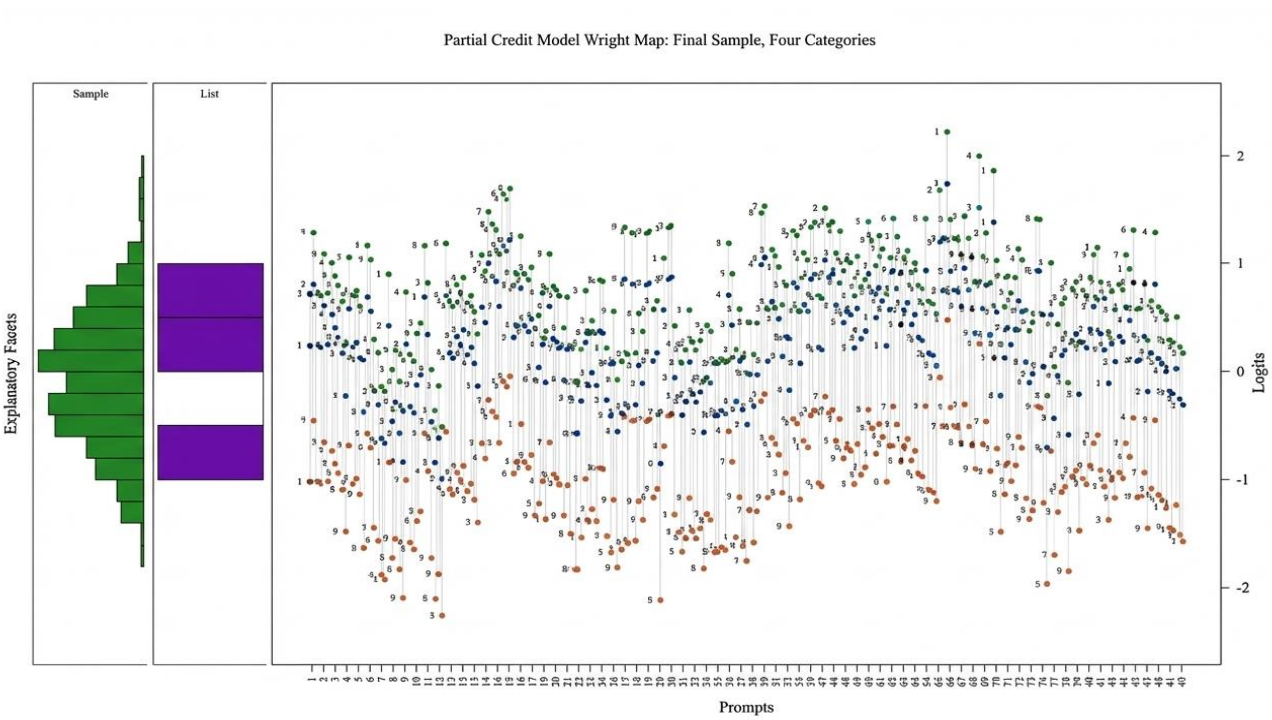

Figure 1. *Wright map for partial credit-many facet Rasch model.*

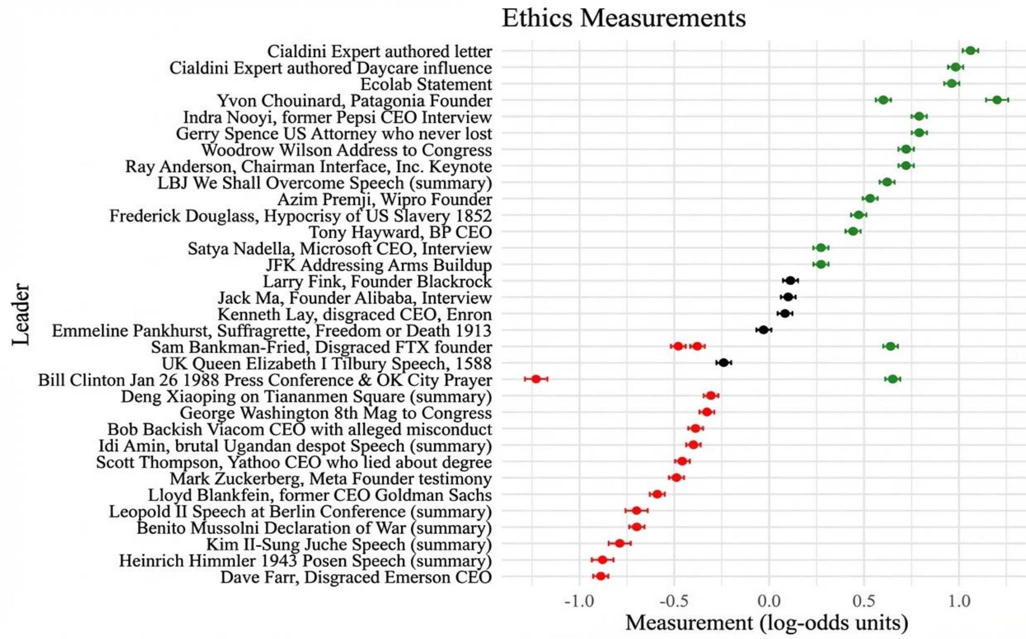

Figure 2. *Ethical measurements of Leader Text and GPT-4 summaries (subset).*

Returning to the Wright Map in Figure 1, the second histogram shows location estimates for LLM raters, which ranged from λ = -0.64 logits for LLM 2 (Wizard LM), to λ = 0.11 logits for LLM 3 (Guanaco), and λ = 0.53 logits for LLM 1 (GPT-3.5). These results suggest that the LLM raters varied with respect to severity.

The large central panel in Figure 1 shows threshold calibrations specific to each prompt. Because we estimated a RS formulation of the MFRM for the final analysis, the distance between thresholds is fixed across prompts, but the thresholds are shifted up or down on the logit scale to reflect overall differences in prompt difficulty. The overall location estimates for prompts ranged from β = -1.25 for the most difficult prompt to β = 1.48 for the easiest prompt.

Explanatory facet calibrations

Table 5 shows detailed results related to the levels of the explanatory facets. For the LLM facet, higher measures correspond to more-severe ratings (i.e., lower ratings); the measures in Table 5 correspond to the locations on the Wright Map, which ranged from λ = -.64 logits for Wizard LM, to λ = .11 logits for Guanaco, and λ = .53 logits for GPT-3.5.

Table 5. *Explanatory facet calibration results: Final sample, rating scale, and model.*

Facet	Level	Observed Average	Fair Average	Measure	*SE*	Infit *MSE*	Outfit *MSE*	Point-Measure Correlation
LLM	GPT-3.5	2.13	1.94	.53	.00	1.06	1.08	.55
	WizardLM	2.90	2.87	-.64	.00	0.85	0.85	.64
	Guanaco	2.37	2.25	.11	.00	1.10	1.15	.49
Sex	Female	2.53	2.37	.04	.00	1.01	1.03	.60
	Male	2.40	2.29	-.04	.00	1.00	1.02	.63
Ethics Level	Low	2.03	2.12	-.27	.00	1.10	1.13	.55
	Medium	2.45	2.40	.08	.00	0.97	1.00	.58
	High	2.60	2.50	.19	.00	0.99	1.00	.61

With respect to location estimates for target ethics level, prompts with predicted low ethics levels according to our theoretical framework had the lowest location (η = -0.28), followed by those with medium ethics levels (η = 0.08), and those with the highest ethics levels (η = 0.20). The ordering of the measures for these prompt classifications is important because it indicates that the LLM raters interpreted the ethics levels for the prompts in the intended direction. Adequate fit statistics for each ethics level supports the interpretation of this ordering.

DISCUSSION and CONCLUSION

Considered together, our results suggest that a measurement approach can be useful to address multiple practical and psychometric challenges related to the use of LLMs in ethical reasoning performance assessment. Although these results should be replicated, they hold considerable promise in several areas. Our results have several implications for research and practice related to ethical persuasion and LLM-based scoring more broadly. We discuss our findings with respect to their implications in the following paragraphs.

Implications for Research and Practice in Ethical Reasoning Assessment

In terms of research on ethical persuasion, our results have implications for the design and implementation of research on measuring ethnical persuasion levels. First, our results demonstrate feasibility with respect to incorporating LLMs into ethical persuasion assessments. Our results suggest that psychometric assessment of ethical persuasion is viable using LLM raters and measurement model analyses. Specifically, our results showed acceptable model-data fit and high levels of precision for measuring ethical persuasion. These results suggest that LLM raters, prompts, settings, and text samples (i.e., examinees) can be defensibly measured with respect to ethical persuasion levels that reflect Cialdini's principles. Importantly, our measures were precise enough to support distinguishing between LLMs with respect to rater severity. We found that GPT3.5 acted as the most severe rater, and WizardLM was the most lenient when scoring ethical persuasion. Despite these differences in severity, ratings from all three LLMs generally adhered to measurement model requirements, making them individually, and together, useful synthetic raters in this context. These results suggest that we can reliably measure a wide range of ethical persuasion texts in an LLM-assisted performance assessment context. For example, text samples from senior leaders, presidents, and prime ministers were scored in ways that we would expect a priori, as shown in Figure 2. This result suggests that it may be possible to use LLMs to evaluate ethical persuasion text in various contexts. For example, it may be possible to use LLM raters to monitor human or LLM-generated text in chatbots or human assessments, to make sure that they remain ethical and generate higher quality feedback before delivering to a user.

Several patterns within our results warrant interpretation. First, GPT-3.5's greater severity (λ = 0.53) compared to WizardLM (λ = -0.64) likely reflects training differences: GPT-3.5's RLHF process may have instilled conservative ethical judgments, while open-source models trained on broader corpora may exhibit greater leniency. This severity difference, while statistically significant, did not compromise measurement quality—all three LLMs exhibited acceptable fit. This suggests ensemble approaches combining multiple LLMs may balance severity biases. These results highlight the importance of combining LLM scoring with a measurement model to control for severity differences to support a meaningful interpretation of LLM rater judgments. Second, the ordering of ethics levels (low: η = -0.28; medium: η = 0.08; high: η = 0.20) provides validity evidence for our theoretical framework. LLMs correctly distinguished developmental progression, supporting Cialdini's hierarchical model. However, the modest separation between medium and high levels suggests these categories may be harder to distinguish in text alone—consistent with MHC theory that higher-order integration is cognitively subtle. Third, the need to collapse our 10-

point scale to four categories reveals an important limitation: LLMs struggle with fine-grained ethical distinctions. This likely reflects both genuine measurement challenges (adjacent MHC stages share substantial overlap) and LLM limitations in detecting nuanced principle integration. Future work should explore whether multimodal assessment (text + audio + video) enables finer discrimination.

Stemming from these strong psychometric findings, our results have implications for practice related to training in ethical persuasion. Specifically, our assessment process supports developmental targeting. By locating learners on the construct map (Table 1), practitioners can calibrate feedback within their Zone of Proximal Development (Vygotsky, 1981)—assigning appropriately challenging tasks between independent and supported performance levels. This addresses a critical gap: most persuasion training lacks developmental precision, risking either trivial exercises for advanced practitioners or overwhelming complexity for novices.

For practitioners, these results enable three applications. First, as a 'flight simulator': professionals can test persuasive messages before high-stakes delivery, receiving immediate developmental feedback calibrated to their MHC stage. Second, for continuous development: past persuasion attempts can be retrospectively analyzed, identifying missed principles or amplifiers to systematically strengthen future performance. Third, for relationship protection: well-intentioned communicators can detect inadvertently unethical language before damaging trust—particularly valuable given that ethical violations disproportionately erode social capital. For organizations, this approach offers proactive risk management. By screening communications before deployment, firms can identify potential manipulators or careless actors whose language patterns suggest ethical deficits—preventing reputational damage and relationship erosion. The measurement precision supports defensible personnel decisions when ethics concerns arise.

Implications for Assessment Research and Practice related to LLMs

Our results also have implications for assessment with LLMs more broadly. Specifically, our results build upon current and emerging research on the use of LLMs as raters, which have generally focused on evaluating writing proficiency and comparisons between LLM and human raters (e.g., Almegren *et al*., 2024; Awidi, 2024; Uyar & Büyükahıska, 2025), connecting these developments to the context of ethical persuasion. Although we focused on ethics and persuasion, the methodology for measurement guardrails is domain-agnostic, and could generalize to a wide range of situations where there is good theory about what low, medium and high looks like in LLMs or multimodal models (e.g. audio, visual samples).

As part of our study design, we explored the degree to which different types of prompt features impact measurement quality. With respect to prompt length, we found that shorter techniques (e.g. using examples and emojis) may be sufficient to avoid the lengthier BARS approaches. This result has important practical implications for prompt development in the context of AI-assisted performance assessment. Similarly, insight related to the impact of prompt features can inform the development of all assessments with LLMs, so that the relative efficacy of prompts is comparable across different samples, LLMs and languages.

Limitations

Our study has some limitations that warrant investigation in future research. Our first limitation reflects measurement concerns related to ethical persuasion. Although we were able to create trustworthy ethics measures from text samples, it is important to recognize that ethical language does not always appropriately reflect Cialdini’s principles. Being ethical is the most basic requirement, but novices may forget or not know about any Cialdini principles. Just because someone does not intentionally misrepresent the truth, contriving it or genuinely believing they are wise does not mean they are not also bungling the use of Cialdini principles, or saying other things that are fiction that they sincerely believe to be fact. Similarly, although two of our gibberish/word salad samples were able to be thrown out from our misfit analyses, one made it through as not deceptive, even though it was objectively nonsensical. For these reasons, it is likely that ethical measurement guardrails are necessary but insufficient to ensure human or LLM textual output quality, and measurement guardrails are also needed for the rest of Cialdini's model, as well as for the facts and evidence that other computer science approaches also address (e.g. ontologies).

This limitation reveals why measurement guardrails are necessary but insufficient. Ethical language is foundational—without it, persuasion becomes manipulation. But ethical language alone doesn't guarantee competent application of Cialdini's principles. A novice can be scrupulously honest yet fail to activate reciprocity, liking, or unity. Conversely, our gibberish sample that passed misfit detection illustrates that current LLMs can mistake nonsense for neutrality. This suggests a two-stage screening process: first, ethical guardrails to prevent manipulation; second, competence assessment across all seven principles to ensure developmental appropriateness.

Another important limitation is related to our sample. We found it hard to secure a sufficiently large number of real human samples that reflect the full range of ethical proficiency that would be ideal, so we synthetically created a large portion of our sample. While early research suggests that GPT-4 is very human-like (Bubeck *et al*., 2023), more evidence is required before we want to conclude that purely engineered samples are sufficiently similar. In particular, we wanted to demonstrate rigorous bias analysis (DPF) with both real human female and synthetic samples, but we did not study other variables of human diversity (e.g. ethnic, ideological, or cultural bias), or combinations of these variables. Researchers may use similar DPF methodology to explore these or other relevant grouping variables in future studies. Further, although our study included global, cross-cultural human examples, we did not explicitly test the fact that some English samples were translations, or summaries of translations and may have reflected idiosyncratic features of GPT-4's ability to generate these summaries, rather than the real nature of the latent attributes in the real leader's native language. At the same time, our current results are conservative, as translation or summary noise should have limited our precision, but did not appear to significantly hinder our measurements of those summaries.

Third, we were primarily interested in whether both closed- and open-source LLMs could provide useful synthetic ratings at all, rather than comparing their relative quality. Although all three LLMs adequately fit the Rasch model, and as an ensemble produced extremely high-quality results, we did not directly compare GPT 3.5, Guanaco and WizardLM. In future studies, researchers may consider using bootstrapping or jackknifing procedures for slightly different configurations of prompts and samples to evaluate whether there are significant quality differences between each of the LLMs, as measured using Rasch quality statistics, as we demonstrated with infit and outfit, along with traditional engineering process capability estimates like CpM and Six Sigma. Relatedly, future research is also needed to more fully explore how the impact of prompt features vary by LLM. Finally, researchers may consider how accuracy-based Rasch models described in Engelhard and Wind (2018) offer additional insight into LLM rating quality with respect to prompt and text characteristics.

Conclusion

Our study combined Cialdini's ethics with measurement theory. We not only demonstrated the reliability and precision of large language models as synthetic raters, but also showed how they can be used to discern ethical nuances in human or AI language, a unique contribution to the burgeoning field of AI safety. As this study pioneers the method of using measurement guardrails for AI systems, it opens the gate to a new era of ethical oversight and quality control in AI and human interactions. While there is much yet to be understood and refined, especially with cross-cultural concerns, these findings give us strong hope, tools, and directions in addressing pressing ethical challenges in AI and human resource development.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Data Availability Statement

Research data are not available for this study.

Informed Consent

No human subjects data were used in this study.

Contribution of Authors

Matt Barney: Investigation, Resources, Visualization, Formal Analysis, and Writing-original draft. **Stefanie A. Wind**: Methodology, Formal Analysis, Visualization, and Writing-original draft. **Vaishak Krishna**: Formal Analysis, Visualization, and Writing-original draft.

Orcid

Matt Barney https://orcid.org/0000-0002-4538-3194
Stefanie A. Wind https://orcid.org/0000-0002-1599-375X
Vaishak Krishna https://orcid.org/0009-0001-1951-9801

REFERENCES

Andrich, D.A. (2018). Rasch rating-scale model. In W.J. van der Linden (Ed.), *Handbook of Item Response Theory* (Vol. 1, pp. 31-50). CRC Press.

Barney, M.F. (2019). The Reciprocal roles of artificial intelligence and industrial-organizational psychology. In R.N. Landers (Ed.), *Cambridge Handbook of Technology and Employee Behavior* (pp. 3-21). Cambridge University Press.

Boyd, R.L., Ashokkumar, A., Seraj, S., & Pennebaker, J.W. (2022). *The development and psychometric properties of LIWC-22*. University of Texas at Austin. https://www.liwc.app/static/documents/LIWC-22%20Manual%20-%20Development%20and%20Psychometrics.pdf

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M.T., & Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4* (Version 5). arXiv. https://doi.org/10.48550/ARXIV.2303.12712

Chan, K.K.Y., Bond, T., & Yan, Z. (2023b). Application of an Automated Essay Scoring engine to English writing assessment using Many-Facet Rasch Measurement. *Language Testing*, *40*(1), 61-85. https://doi.org/10.1177/02655322221076025

Chia, Y.K., Hong, P., Bing, L., & Poria, S. (2023). *INSTRUCTEVAL: Towards Holistic Evaluation of Instruction-Tuned Large Language Models* (No. arXiv:2306.04757). arXiv. https://doi.org/10.48550/arXiv.2306.04757

Cialdini, R.B. (2009). *Influence: Science and practice* (5. ed., internat. ed). Pearson Education [u.a.].

Cialdini, R.B. (2014). *Influence*. HarperCollins e-Books.

Cialdini, R.B. (2016). *Pre-Suasion: A revolutionary way to influence and persuade*. Simon & Schuster.

Commons, M.L. (2007). Introduction to the model of hierarchical complexity. *Behavioral Development Bulletin*, *13*(1), 1–6. https://doi.org/10.1037/h0100493 de Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach* (2004th edition). Springer.

Engelhard, G., & Myford, C.M. (2003). *Monitoring faculty consultant performance in the Advanced Placement English Literature and Composition Program with a many-faceted Rasch model*. College Entrance Examination Board.

Engelhard, G., & Wind, S.A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Taylor & Francis.

Engelhard Jr, G., & Wang, J. (2020). *Rasch Models for Solving Measurement Problems* (Vol. 187). Sage. https://us.sagepub.com/en-us/nam/rasch-models-for-solving-measurement-problems/book267292

Fischer, K.W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review*, *87*(6), 477–531. https://doi.org/10.1037/0033-295X.87.6.477

Flodén, J. (2025). Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British Educational Research Journal*, *51*(1), 201–224. https://doi.org/10.1002/berj.4069

He, K., Pu, N., Lao, M., & Lew, M.S. (2023). Few-shot and meta-learning methods for image understanding: A survey. *International Journal of Multimedia Information Retrieval*, *12*(2), 14. https://doi.org/10.1007/s13735-023-00279-4

Kennedy, C.J., Bacon, G., Sahn, A., & Vacano, C. von. (2020). *Constructing interval variables via faceted Rasch measurement and multitask deep learning: A hate speech application* (No. arXiv:2009.10277). arXiv. https://doi.org/10.48550/arXiv.2009.10277

Kim, H., Baghestani, S., Yin, S., Karatay, Y., Kurt, S., Beck, J., & Karatay, L. (2024). ChatGPT for Writing Evaluation: Examining the Accuracy and Reliability of AI-Generated Scores Compared to Human Raters. In C. Chapelle, G. Beckett, & J. Ranalli (Eds.), *Exploring AI in Applied Linguistics* (pp. 73–95). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2024.154.06

Linacre, J.M. (1989). *Many-Facet Rasch Measurement*. MESA Press.

Linacre, J.M. (2020). *A user's guide to FACETS: Rasch-model computer programs* (Version 3.83.4) [Computer software]. winsteps.com. http://www.winsteps.com/manuals.htm

Masters, G.N. (2018). Partial credit model. In W. J. van der Linden (Ed.), *Handbook of Item Response Theory* (Vol. 1, pp. 109-126). CRC Press.

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

Page, E.B. (1966). The imminence of grading essays by computer. *The Phi Delta Kappan*, *47*(5), 238-243.

Pfau, A., Polio, C., & Xu, Y. (2023). Exploring the potential of ChatGPT in assessing L2 writing accuracy for research purposes. *Research Methods in Applied Linguistics*, *2*(3), 100083. https://doi.org/10.1016/j.rmal.2023.100083

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests* (Expanded edition, 1980). University of Chicago Press.

Ramesh, D., & Sanampudi, S.K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, *55*(3), 2495–2527. https://doi.org/10.1007/s10462-021-10068-2

Reckase, M.D. (1979). Unifactor Latent Trait Models Applied to Multifactor Tests: Results and Implications. *Journal of Educational and Behavioral Statistics*, *4*(3), 207-230. https://doi.org/10.3102/10769986004003207

Sachdeva, P.S., Barreto, R., Von Vacano, C., & Kennedy, C.J. (2022). Assessing Annotator Identity Sensitivity via Item Response Theory: A Case Study in a Hate Speech Corpus. *2022 ACM Conference on Fairness Accountability and Transparency*, 1585–1603. https://doi.org/10.1145/3531146.3533216

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). *Towards a standard for identifying and managing bias in artificial intelligence* (No. NIST SP 1270; p. NIST SP 1270). National Institute of Standards and Technology (U.S.). https://doi.org/10.6028/NIST.SP.1270

Seßler, K., Fürstenberg, M., Bühler, B., & Kasneci, E. (2025). Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring. *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 462-472. https://doi.org/10.1145/3706468.3706527

Smith, R.M. (2004). Fit analysis in latent trait models. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch Measurement* (pp. 73–92). JAM Press.

Solaiman, I., Talat, Z., Agnew, W., Ahmad, L., Baker, D., Blodgett, S.L., Chen, C., III, H.D., Dodge, J., Duan, I., Evans, E., Friedrich, F., Ghosh, A., Gohar, U., Hooker, S., Jernite, Y., Kalluri, R., Lusoli, A., Leidinger, A., … Subramonian, A. (2024). *Evaluating the Social Impact of Generative AI Systems in Systems and Society* (No. arXiv:2306.05949). arXiv. https://doi.org/10.48550/arXiv.2306.05949

Sun, X., Gu, J., & Sun, H. (2021). Research progress of zero-shot learning. *Applied Intelligence*, *51*(6), 3600–3614. https://doi.org/10.1007/s10489-020-02075-7

Vygotsky, L.S., & Cole, M. (1981). *Mind in society: The development of higher psychological processes* (Nachdr.). Harvard Univ. Press.

Walker, A.A., & Wind, S.A. (2020). Identifying Misfitting Achievement Estimates in Performance Assessments: An Illustration Using Rasch and Mokken Scale Analyses. *International Journal of Testing*, *20*(3), 231–251. https://doi.org/10.1080/15305058.2019.1673758

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* (Version 6). arXiv. https://doi.org/10.48550/ARXIV.2201.11903

Wetzler, E.L., Cassidy, K.S., Jones, M.J., Frazier, C.R., Korbut, N.A., Sims, C.M., Bowen, S.S., & Wood, M. (2024). Grading the graders: Comparing generative AI and human assessment in essay evaluation. *Teaching of Psychology*, 00986283241282696. https://doi.org/10.1177/00986283241282696

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Taylor & Francis.

Wilson, M. (2023). *Constructing measures: An item response modeling approach* (Second edition). Routledge.

Wu, M., & Adams, R.J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement*, *14*(4), 339–355.

Yamashita, T. (2024). An application of many-facet Rasch measurement to evaluate automated essay scoring: A case of ChatGPT-4.0. *Research Methods in Applied Linguistics*, *3*(3), 100133. https://doi.org/10.1016/j.rmal.2024.100133

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T.L., Cao, Y., & Narasimhan, K. (2023). *Tree of Thoughts: Deliberate Problem Solving with Large Language Models* (No. arXiv:2305.10601). arXiv. https://doi.org/10.48550/arXiv.2305.10601

Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology*, *56*(1), 150–166. https://doi.org/10.1111/bjet.13494

Yun, J. (2023). Meta-analysis of inter-rater agreement and discrepancy between human and automated English essay scoring. *English Teaching*, *78*(3), 105-124. https://doi.org/10.15858/engtea.78.3.202309.105

Zhang, Y., Jin, R., & Zhou, Z.-H. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*, *1*(1-4), 43-52. https://doi.org/10.1007/s13042-010-0001-0

APPENDICES

To gain confidence that our hypothesized approach would be worth testing in a formal study, we created a small proof-of-concept pilot to study both Cialdini's approach to ethics and one of his principles of persuasion, reciprocity, reported elsewhere (Barney, Under Review).

Cialdini's framework for ethics provides a superstructure for a wide range of scientific models related to ethics, and informed our creation of prompts that we hypothesized could detect persuasive textual samples that are a) True (factual), b) Intrinsic, or natural to the situation (not contrived or exaggerated), and c) Wise (beneficial for all stakeholders). This provided the superstructure for a detailed construct map for unethical behavior, to account for gradations of ethical proficiency, provide traceability back to theory and evidence, and make sure our instrument had a good chance at having a sufficient amount of information along the full continuum.

We conceptualized ethics as a micro-construct map of Cialdini's model, and because we wanted to understand what prompt engineering methods were most effective, we intentionally authored a wide variety of prompts not only at different levels of ethics, using a variety of prompt authoring methods from computer science. These include:

- Zero Shot learning: enables LLMs to infer the correct output for tasks on which they have not been explicitly trained (Sun, Gu & Sun, 2021)
- Few Shot learning: training LLMs on a very small amount of example data, asking them to generalize to novel examples (He, Pu, Lao *et al*, 2023)
- Chain of Thought prompts: written to guide the LLM to articulate intermediate reasoning steps that improves performance on complex tasks. (Wei, Wang, Shen, Eban, Zhao & Ouyang, 2022)
- Bag of Words prompts: written based on pre-LLM research showing statistical representations where text is a set of unordered words, disregarding grammar and word order, and often used with Classical Test Theory methods (Boyd, Ashokkumar, Seraj, & Pennebaker, 2022; Zhang, Jin & Zhou, 2010)

In addition, because psychometrics has found improvements with human raters with the use of behaviorally-anchored rating scales (BARS) and examples (e.g. of gradations of true, natural, and wise), we included BARS and examples as hypothesized improvements. For the prompts that used BARS we hypothesized that we could use the essence of Wilson's (2023) Construct Map approach to approximate Cialdini's levels of ethics at each level of ethics proficiency ordinally on the 1-10 rating scale, within each prompt as well as between prompts.

The first author was certified by Cialdini to teach his methods and used that knowledge in prior failed attempts to create an artificially intelligent approach to assessing and coaching persuasion using earlier machine learning approaches that pre-date Large Language Models (Barney, 2018). In this study, he used this knowledge to both manually author prompts, and prompt-engineer others using GPT-4. While authoring, he scrutinized the literature from various schools of philosophy, cultural anthropology, theology, sociology, computational linguistics, social psychology, industrial-organizational psychology, political science and economics to write 110 total pilot prompts focused on ethics to be tested. Table 3 shows sample prompts for each facet of Cialdini's framework.

The original pilot used GPT3.5 Turbo to make ratings in order to eliminate the higher cost of using GPT-4. Using GPT for Sheets, we submitted each prompt, and a vector of samples either intentionally engineered to reflect the scientific conception of Cialdini's framework with GPT-4, or the author selected a real persuasive example from his archive of expert-written persuasive appeals. We also included control samples with gibberish, or "persuasion word salad" that included persuasive words but were nonsensical.

To see if temperature affected the ability of the LLM to emulate a good rater, we tested four different temperature settings across the full range of possible levels: 0, 0.25, 0.5, & 1.0. We also tested whether human or GPT3.5 generated samples were both equally useful, as well as a set of prompts. The Many Facet Rasch Model used in the Pilot was:

$$log\left(\frac{P(X_{nijk} = p)}{P(X_{nijk} = p - 1)}\right) = \theta_n - \delta_i - \alpha_j - \beta_k - \tau_p$$

where:

θ_n is the location of the GPT3.5 Turbo's temperature settings

δ_i is the location of the sample (person)

α_j is effect of human authored or GPT-4 created sample

β_k is the location of the prompt

τ_p is the Rasch-Andrich threshold parameter for category p of the 10 point rating scale

Using Facets version 3.85.1 (Linacre, 2022), we were able to create an instrument successfully using Joint Maximum Likelihood Estimation, and Linacre's Rating Scale Model. A principle components analysis showed that 83.4% of the variance were accounted for by the pilot measure, suggesting we successfully produced a unidimensional measure, after removing misfitting prompts. While all temperature settings fit the Rasch Model's quality standards, Appendix 1 shows a Wright Map produced by Facets that also visually shows that all of the temperature settings were of the same level of severity/leniency, suggesting the different temperatures didn't add new information about low or high prompts or samples. The same Wright Map also shows other facets on the same instrument: persuasive samples, sample sources (GPT-4 generated higher difficulty samples than the human samples), the prompts (labeled items in the Wright Map), all on the same logit scale (far left labeled "Measr"); transformed back to the original rating scale (Far right labeled "AGREE").

Because there were no significant differences in the usefulness of different temperature ratings (set at 0, 0.25, 0.5 and 1), we dropped those from the full study. Even though the analysis showed that the human-generated samples were slightly lower than those generated by GPT-4 in this sample, they were equally useful in helping to estimate the location of the other facets. Wright Maps are powerful because they allow the analyst to see all measured facets in one frame of reference, and quickly see the lack of difference in temperature and a small difference in the source of the samples while clearly seeing large variations in samples, and prompts.

Interestingly, GPT3.5 turbo had a hard time discerning meaningful distinctions from all 10 of the rating options in our scale, as is very common with real human raters in traditional use of the Many Facet Rasch Model in psychology. While GPT-3.5 could use most levels like an expert human, it failed to discern the difference between a rating of 4 and 5, and between a 6 and 7 on the 10 point scale. Consequently, we combined these categories and the entire instrument fit the Rasch model well.

When considering these results in the context of current quality standards for high-stakes human tests such as physician credentialing, we consider the pilot a success. Appendix 2 shows these comparisons. Separation and strata are two ways to estimate how many unique units can be discerned given the measurement uncertainty present in an instrument. When considering standards for physicians, our experimental pilot measures of prompt quality showed 8.6 to 11.9 unique levels that can be discerned between them; and crucially an extremely high 32.3 to 43.5 gradations of ethics in the samples, which is typically the focus of an assessment. As the table in Appendix 2 shows, this is more than 4 times more precise than the Clinical Pathology test results reported by O'Neill and used with physicians (1997). This successful pilot test justified a complete study.

Appendix 1. *Wright map of all facets in the pilot study.*

```
+---------------------------------------------------------------------------------------------+
|Measr|+Algorithm                                        |+Sample        |+Source|+Item  |AGREE|
|-----+-------------------------------------------------+---------------+-------+-------+-----|
|   5 +                                                 +               +       +       +(10) |
|     |                                                 |               |       |       |     |
|     |                                                 |               |       |       |     |
|     |                                                 |               |       |       |     |
|     |                                                 | GPT4 HRLE     |       |       |     |
|   4 +                                                 +               +       +       +     |
|     |                                                 |               |       |       |     |
|     |                                                 |               |       |       |     |
|     |                                                 | GPT4 HRHE     |       |       |     |
|     |                                                 |               |       |       |     |
|   3 +                                                 + Turbo Frank   +       +       + --- |
|     |                                                 |               |       |       |     |
|     |                                                 |               |       |       |     |
|     |                                                 | GPT4 MRHE     |       |       |     |
|     |                                                 |               |       |       |     |
|   2 +                                                 +               +       +       +  8  |
|     |                                                 | Frank's Lady  |       | *     |     |
|     |                                                 |               |       |       |     |
|     |                                                 | Daycare       |       |       |     |
|     |                                                 |               |       | *     |     |
|   1 +                                                 +               +       + **    + --- |
|     |                                                 |               |       | ***   |     |
|     |                                                 | Merck Food    | Human |       |     |
|     |                                                 | CEO           |       | *     |     |
|     | T0 g3.5 turbo   T0.5 3.5 turbo  T1 3.5 turbo    |               |       | ***** |  4  |
*   0 *                                                 *               *       *       *     *
|     |                                                 | Spence        |       | *     |     |
|     | GPT-4                                           |               |       | **    |     |
|     |                                                 | GPT4 MRLE     | GPT4  | *     | --- |
|     |                                                 |               |       | *     |     |
|  -1 +                                                 +               +       + *     +     |
|     |                                                 |               |       |       |     |
|     |                                                 |               |       |       |     |
|     |                                                 |               |       | *     |     |
|     |                                                 |               |       | *     |  2  |
|  -2 +                                                 +               +       + *     +     |
|     |                                                 |               |       |       |     |
|     |                                                 |               |       |       |     |
|     |                                                 |               |       |       |     |
|     |                                                 |               |       |       |     |
|  -3 +                                                 +               +       +       +     |
|     |                                                 |               |       |       |     |
|     |                                                 | GPT4 LRLE     |       |       | --- |
|     |                                                 |               |       |       |     |
|     |                                                 |               |       |       |     |
|  -4 +                                                 +               +       +       + (1) |
|-----+-------------------------------------------------+---------------+-------+-------+-----|
|Measr|+Algorithm                                        |+Sample        |+Source| * = 1 |AGREE|
+---------------------------------------------------------------------------------------------+
```

Appendix 2. *Pilot results compared with high-stakes tests.*

	Separation†	Strata¹	Reliability
Optometry Council of Australia & New Zealand - Items (Backhouse *et al*., 2021)	2.54 to 4.06	3.69 to 5.42	0.82 to 0.93
Objective Structured Clinical Examination (Medical Residents) - Items (Sloan *et al*., 1996)	2.33 to 3.54	3.44 to 4.72	0.78 to 0.91
American Board of Physical Medicine and Rehabilitation Part 1 Certification Examination Items - Items (Raddatz *et al*., 2017)	4.06 to 7.0	5.42 to 9.33	0.93 to 0.98
Physician Transfusion Medicine Knowledge - Items (Haspel *et al*., 2014)	2.48 to 2.58	3.64 to 3.77	0.79 to 0.80
RN Comprehensive Predictor - Items (ATI, 2020)	1.94 to 3.16	3.24 to 4.21	0.70 to 0.96
Clinical Pathology - physicians (O'Neill *et al*., 1997)	2.67	3.89	0.81
Clinical Pathology - tasks (O'Neill *et al*., 1997)	7.07	9.43	0.99
Clinical Pathology MD Judge severity/leniency (O'Neill *et al*., 1997)	5.57	7.43	0.97
Cialdini Influence Assessment‡	5.27 to 7.23	7.35 to 9.97	0.97 to 0.98
Pilot - Items (prompts)	*8.6*	*11.9*	*0.99*
Pilot - Persons (samples)	*32.3*	*43.5*	*>0.99*

† Estimated from reported reliability in cited studies

‡ First author previously helped create an Automatic Item Generation version of Cialdini's Influence IQ Test, now called the Cialdini Influence Assessment that is owned by http://cialdini.com