Behavior Research Methods (2026) 58:150
https://doi.org/10.3758/s13428-026-03016-x

ORIGINAL MANUSCRIPT

# Measuring intellectual humility through situated behavior: An alternative to dispositional self-reports

Maksim Rudnev[1] · Ana Lucia Rodriguez de la Rosa[2] · Ryan Barrett[2] · Nicholas A. Christakis[2,4] · Igor Grossmann[1,3]

Received: 1 July 2025 / Accepted: 30 March 2026 / Published online: 4 May 2026


## Abstract

Although psychological theory views behavior as an interaction between person and situation, the measurement of metacognition often relies on situation-free, abstract self-views. To align measurement with interactionist frameworks, we proposed a situated approach, using intellectual humility (IH)—recognizing one's limits and fallibility—as a proof-of-concept. We validated this approach across diverse populations: English- and Spanish-speaking North Americans ($N = 633$) and adults from 136 rural Honduran villages ($N = 2567$). Rather than rating abstract tendencies, participants reconstructed three recent disagreements (freely chosen, wrong, and right) and reported specific behaviors via branching binary probes. This method demonstrated cross-cultural coherence of the IH construct while capturing substantial situational variability. Notably, 74–81% of variance occurred *within* persons: IH expression fluctuated significantly based on epistemic context (being wrong vs. right) and social dynamics (partner status). These situational effects also fully explained gender differences in the Honduran sample. The situated approach showed efficiency outside Western, educated contexts and helps overcome the *humility paradox*—wherein the least intellectually humble overclaim their humility. We discuss four principles for aligning measurement with theory—contextual specificity, sampling from actual experiences via event reconstruction, accessibility across diverse populations via branching probes, and modeling within-person variability—offering a framework for assessing metacognitive and self-regulatory constructs beyond the constraints of static dispositional measures.

**Keywords** Cross-cultural measurement · Metacognition · Multilevel constructs · Dispositional bias

## Introduction

Psychological science has long embraced an interactionist view of behavior (Funder, 2006; Lewin, 1936; Mischel & Shoda, 1995). Under this framework, human action reflects a dynamic interplay between personal dispositions and situational demands; the same underlying disposition may manifest differently across people, and crucially, within the same person across varying contexts. This interactionist framework is foundational to modern theories of cognition, affect, motivation, and particularly metacognition—the processes by which people monitor and regulate their thinking to fit situational demands. Yet, despite theoretical consensus, dominant measurement approaches often treat metacognitive characteristics as stable traits (for a review, see Ballantyne, 2023; Fleeson & Jayawickreme, 2021). By assuming that underlying traits drive behavior consistently across contexts, these dispositional measures disconnect from nuanced interactionist theories, masking substantial intra-individual variability (e.g., Funder, 2006). This problem compounds when researchers rely on global self-assessments. Due to their lack of specificity, such measures invite motivated reasoning (Kunda, 1990), social desirability bias (Paulhus & Reid, 1991), and self-enhancement (Sedikides & Gregg, 2008) and introduce measurement error that undermines construct validity.

### The case of intellectual humility

The gap between interactionist theory and dispositional measurement is particularly acute in constructs that are inherently

✉ Maksim Rudnev
maksim.rudnev@gmail.com

1 Department of Psychology, University of Waterloo, Waterloo, ON, Canada

2 Department of Sociology, Yale University, New Haven, CT, USA

3 African Institute for Epistemology and Philosophy of Science, University of Johannesburg, Johannesburg, South Africa

4 Human Nature Lab, Yale University, New Haven, CT, USA

context-dependent (e.g., Costello et al., 2023), such as intellectual humility (IH). Defined as the recognition of one's intellectual limitations (Porter et al., 2022b; Whitcomb et al., 2017), IH involves admitting knowledge gaps, revising views, and engaging respectfully with alternative perspectives (Koetke et al., 2022; Krumrei-Mancuso et al., 2020; Whitcomb et al., 2017). While theories of IH emphasize that it emerges from the interaction between dispositions and the environment (e.g., Hoyle et al., 2016; Krumrei-Mancuso et al., 2025; Leary et al., 2017; see Porter et al., 2022b, for a review), dominant measurement practices have lagged behind. Whether an individual demonstrates IH depends on the social context (e.g., Dorfman et al., 2022; Grossmann & Kross, 2014; Kube & Rozenkrantz, 2021; Smith et al., 2025; Zachry et al., 2018), the specific subject matter (e.g., Dunning, 2011), and may be related to the current affective states (e.g., Schwarz & Clore, 1983). Yet, typical measurement approaches to IH rely on abstract, trait-level self-evaluations (e.g., "I am willing to change my position on an important issue in the face of good reasons" – Krumrei-Mancuso & Rouse, 2016). Though useful for mapping broad individual differences in self-views and pragmatic in terms of study length, this approach has several blind spots that we will review next.

### Cognitive accessibility

Dispositional measures rest on two precarious assumptions: that participants can accurately access their general IH tendencies and that situational variability is negligible (also see Smith et al., 2025). However, the cognitive demands of accurately reporting general tendencies are immense. An unbiased response requires a participant to: (1) recall all instances of facing conflicting evidence; (2) evaluate their IH in each instance; and (3) aggregate these disparate data points into a single summary score.

This "mental calculus" exceeds typical memory capacities. Consequently, participants likely rely on heuristics (e.g., "What answers seem socially appropriate?"), leading to biases such as domain imbalance (e.g., weighing rare political discussions heavily while ignoring daily interactions) and recall bias (e.g., favoring events in which they successfully changed their minds). For instance, an individual may demonstrate high IH in culinary contexts but display arrogance in political debates. Although the interactionist framework acknowledges the situational variability of IH in theory, trait-level scales collapse these nuances, in practice failing to capture an important part of the IH variance.

### The humility paradox

Beyond cognitive accessibility and assumption of stability, self-evaluation of trait humility faces a fundamental validity threat we term the "humility paradox." Research in metacognition consistently shows that those with the lowest competence often exhibit the highest confidence (Dunning, 2011; Fleming & Lau, 2014; Metcalfe & Finn, 2013). A similar logic applies here: intellectually arrogant individuals may sincerely believe they are open to revising their beliefs and admit their limits, while those who are genuinely intellectually humble (and thus hyper-aware of their fallibility) may underestimate their virtue (Alfano et al., 2017; Porter et al., 2022b).

This process creates a measurement artifact in which the highest scorers on self-reported humility scales may actually be the least humble. Intellectually humble individuals may be well-calibrated due to their receptivity to feedback; conversely, individuals low in IH are, by definition, likely miscalibrated due to a resistance to recognizing their own limitations. Consequently, high scores on traditional self-reports are often unreliable indicators of the construct they intend to measure.

### Cross-cultural transferability

Current abstract assessment tools reflect the abstract-analytical epistemological traditions of the Global North (Nisbett et al., 2001). As a result, existing IH scales have been validated predominantly within WEIRD (Western, Educated, Industrialized, Rich, and Democratic) populations (Henrich et al., 2010). Given documented cross-cultural differences in cognitive styles (Masuda & Nisbett, 2001; Varnum et al., 2010), reliance on abstract assessment threatens cross-cultural transferability (Grossmann & Na, 2014). Theoretically, the expression and elicitation of IH vary significantly across cultures, social classes, and relational structures (Brienza & Grossmann, 2017; Fiske, 1992; Markus & Kitayama, 1991).

Furthermore, the dominant reliance on numerical rating scales presumes a level of numeracy and survey familiarity that is not universal. Anthropological accounts demonstrate that the format of the questions typically used for psychological measurement might not be appropriate in some cultures (e.g., tending towards "holistic" rather than "analytic" cognitive styles; Nisbett et al., 2001; Poortinga & Van de Vijver, 2004). This methodological bias risks excluding underrepresented populations, suggesting that current tools may lack measurement invariance across diverse groups.

### Challenges of hypothetical in-lab reports

One might argue that situational variability in IH can be captured via laboratory simulations or hypothetical vignettes (e.g., Grossmann & Kross, 2014; Staudinger & Baltes, 1996). However, hypothetical reports present distinct validity challenges. First, standardized scenarios often lack

cross-cultural equivalence; a specific vignette may not be understood or valued equally across contexts (e.g., Hruschka et al., 2018). Second, distinct gaps exist between *competence* (hypothetical capacity) and *performance* (real-world behavior). As seen in moral psychology, abstract questions often fail to predict real-world decisions because they strip away the contextual nuance and affective pressure of lived experience (Bostyn et al., 2018). Finally, by standardizing situations across all participants, one inherently fails to capture naturalistic within-person variability in IH, for example, due to changes in social-ecological contexts.

## A behavioral alternative: Principles for situated measurement

In summary, existing instruments treat IH as a static disposition measured via abstract self-evaluation, introducing recall bias, limited introspective access, and the humility paradox. Moreover, to our knowledge, none of the existing instruments are sensitive to the cultural or socio-economic nuances, such as differences in cognitive styles or numeracy. To address these validity threats and accommodate cross-cultural diversity, we propose four design principles for advancing the measurement of metacognitive tendencies (summarized in Table 1).

1. *Contextual specificity*. We assess *behavior in concrete situations* rather than abstract tendencies. Assessing IH-related behaviors in specific situations allows for clarifying the boundary between situational and trait expressions of a construct (e.g., Fleeson & Jayawickreme, 2021), moving beyond the default assumption of cross-situational consistency.
2. *Event reconstruction*. We sample from people's *actual experiences* rather than hypothetical scenarios. Building on the event reconstruction methods developed for capturing accurate recall of feelings and emotional states (Kahneman et al., 2004; Schwarz & Oyserman, 2001), a recall process focuses on specific, recent events. This approach recruits different memory systems than evaluating general traits, thereby reducing cognitive demand and self-presentation bias (Brienza et al., 2018; Grossmann, 2017; Porter et al., 2022b; also see Schwarz & Oyserman, 2001).
3. *Accessibility optimization*. To increase accessibility beyond WEIRD populations and reduce the cognitive burden of abstract rating scales, we propose to decompose items into a series of simple binary prompts (i.e., a branching or binary-chain format). This interview-style approach reduces ambiguity, accommodates lower numeracy levels, and improves response quality (Hilbert et al., 2016; Krosnick, 1991).

**Table 1** Principle-based considerations and proposed solutions for assessing intellectual humility

| | Principle | Solution | Example | Methodological benefits |
|---|---|---|---|---|
| 1 | Contextual specificity | Assess behavior in construct-relevant contexts | The context is constrained to "a recent difference in opinion" | Ensures general consistency of recalled situations across situations and participants, affording comparability of the upcoming responses |
| 2 | Event reconstruction | Prompt to recall of a recent experience via the event reconstruction protocol | "Please recall your most recent disagreement. Who was it with? When did it happen?" | - Captures real-life behaviors (ecological validity); attenuates memory biases; reduces reliance on semantic self-views<br>- Attenuation of memory biases and help with recall accuracy |
| 3 | Accessibility optimization | Probe a set of construct-relevant behaviors during the recalled event on an accessible scale | Inquiry about six IH-relevant behaviors via *binary-chain prompting*: start with a manifest question (did or did not do it), followed by the degree question (a little vs. very much) | Attenuates cognitive load and overcomes concerns with numeracy in less-educated populations |
| 4 | Modeling variability | Sample responses across a diverse range of episodic states | Participants recall three distinct events: one "freely chosen," one where "you were wrong," and one where "you were right." | Cost-effective measure of cross-situational variability, distinguishing between trait and state without longitudinal delays |

4. *Modeling variability*. Finally, we model within- and between-person variability simultaneously. By sampling responses across distinct situations, we capture the "density profile" of the trait—consistent with Whole Trait Theory (Fleeson & Jayawickreme, 2021). This multi-situation sampling allows for an idiographic analysis of how an individual varies across contexts while maintaining the ability to derive nomothetic trait-level estimates (e.g., Grossmann et al., 2025; Molenaar, 2004).

Although we apply these principles to the construct of intellectual humility, we suggest that they are applicable in a wide range of metacognitive and self-regulatory constructs. These principles can be further integrated with best practices for cross-cultural design, including cognitive interviewing to ensure stimulus relevance (Lenzner et al., 2016) and the inclusion of reverse-coded items to model acquiescence response styles—a critical source of construct-independent variance in cross-cultural research (Billiet & McClendon, 2000; Rivera-Garrido et al., 2022).

### Research overview

We tested this situated approach to intellectual humility by developing and validating a measurement protocol across linguistically and socio-culturally distinct populations (see Fig. 1). We focused on a context where IH is most critical: interpersonal disagreements. When people disagree, they face a choice—defend their position or consider they might be wrong. Together, these moments reveal intellectual humility in action. To capture within-person variability, we used an autobiographical experience sampling design. Participants recalled three recent disagreements using a stratified approach to ensure variance: one freely chosen disagreement, one instance in which they thought they were wrong, and one in which they thought they were right. We hypothesized that situations in which an individual recognizes they were wrong would serve as stronger situational triggers for the metacognitive core of IH—awareness of fallibility—thereby eliciting higher IH than situations in which they were right. This design encouraged participants to recall a disagreement organically (via the "freely chosen" prompt) while ensuring coverage of diverse epistemic outcomes (i.e., being right or wrong). For each disagreement, participants responded to a six-item instrument using a branching format (binary presence/absence check, followed by intensity), designed to minimize cognitive load and maximize translatability. After developing the blueprint of the novel instrument, we proceeded with systematic pre-testing involving cognitive interviews, repeated adjustments, and validity checks, which was followed by the administration in a large representative sample of the population of Copán, Honduras.

## General methods

Our development and validation of this tool proceeded in five steps (see Fig. 1). We developed and validated the inventory through a four-study process designed for cross-cultural application. We first generated a range of candidate stimuli based on prior theory and in consultation with an expert panel. In Study 1, we conducted 30 cognitive interviews in rural Honduras to iteratively probe comprehensibility and contextual fit of the candidate scenario and items. Based on these insights, in Study 2, we evaluated the psychometric properties of the situated instrument on an online, English-speaking US sample. Via this step, we sought to evaluate inventory's dimensionality and its nomological network—i.e., discriminant and convergent validity against established individual-difference measures. In Study 3, we then replicated these psychometric findings with an online, Spanish-speaking US sample. Finally, in Study 4, we administered the inventory through in-person interviews across 136 rural Honduran villages—a low-literacy, resource-limited context, affording a comparison of model fit and cross-situational variability with the online samples from the US. For quantitative analyses, we followed empirically derived estimates of effect sizes in social-personality psychology (Gignac & Szodorai, 2016), treating $r = .10$, 20, and.30 as small, medium, and large effects, respectively. Further, we corrected correlations' $p$ values for multiple comparisons using Benjamini–Hochberg's false discovery rate approach (Murray & Blume, 2021).

Statistical models were run using *Mplus* software, version 8.10 (Muthén & Muthén, 2022), and post-processed with *R* (R Core Team, 2023). Supplementary Information (SI) is available at the OSF directory https://osf.io/wj3fe.

## Study 1

Following the existing approaches (Hazlett, 2012; Hoyle et al., 2016; Krumrei-Mancuso & Rouse, 2016; Leary et al., 2017), we developed a working definition of IH. We further outlined several candidates for locally relevant situations, in which IH is most likely to manifest—i.e., situations involving disagreements and differences of opinion—with continuous adjustment based on feedback from cognitive interview participants and local expert team members. The goal was to ensure the situations were culturally relevant to a low-educated Honduran context, which was the final target of our study.

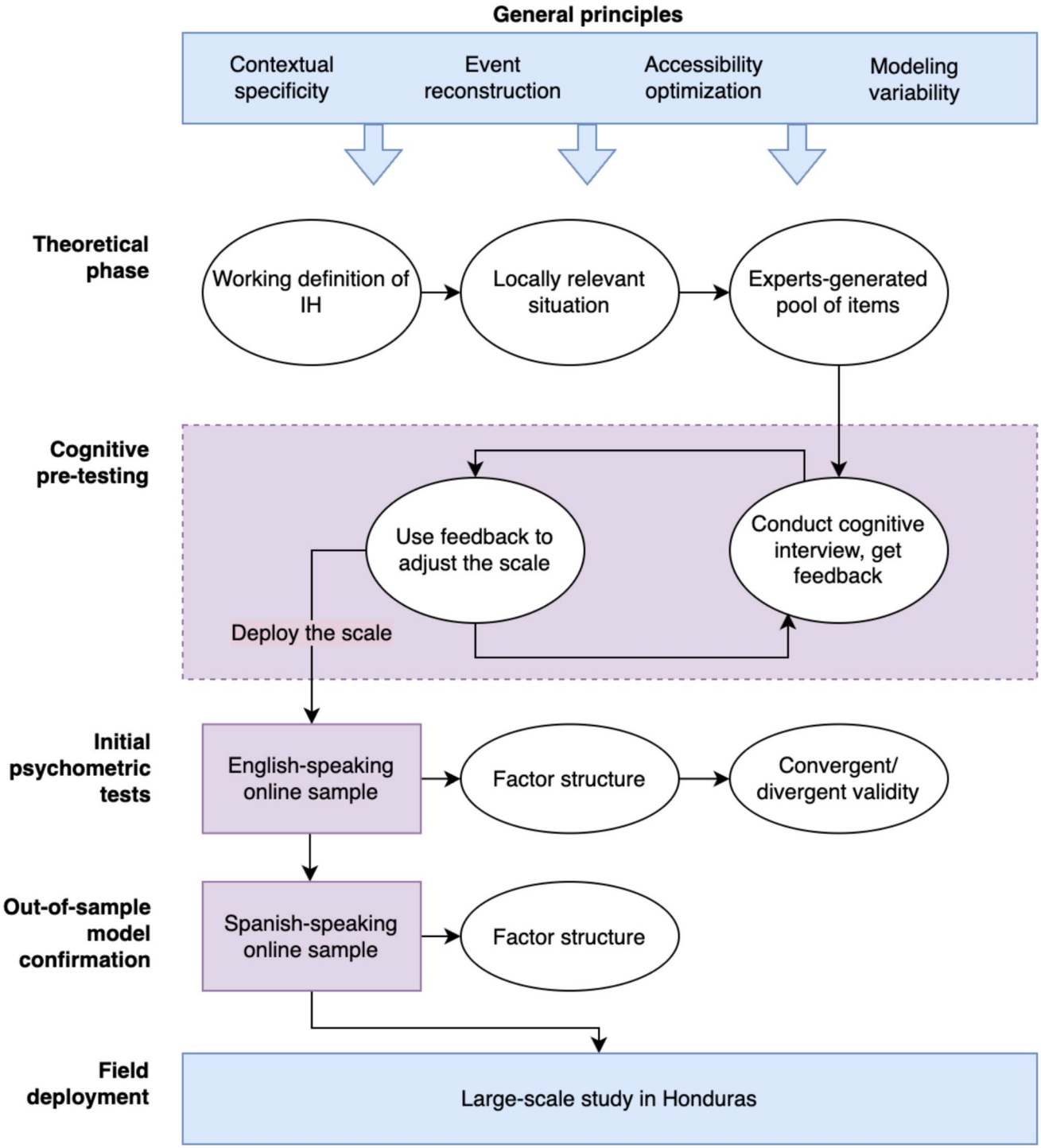


**Fig. 1** Study flow across theoretical, qualitative, and quantitative phases of the project

## Methods

Our group of experts generated several sets of stimuli, depicting metacognitive strategies consistent with existing scholarship on IH (Brienza et al., 2018; Krumrei-Mancuso & Rouse, 2016; Leary et al., 2017; also see Porter et al., 2022b, for review). In consultation with the team of local experts, we first identified items from existing IH scales that were relevant to the Honduran social context. Then, we conducted thirty cognitive interviews (Beatty & Willis, 2007; Schwarz & Oyserman, 2001) among individuals from our target population of rural Hondurans, iteratively

adjusting the item wordings, the situation descriptions, and the response options after each interview until the set of criteria was satisfied (see OSF for the list of pilot activities). The criteria included both the general principles outlined in Table 1, along with pragmatic considerations of local relevance and cost-effectiveness.

### Participants

Participants were recruited from the rural villages in the Copán Department, Honduras, which was the main field study site for the present project. All participants were part of a previously established panel (Airoldi & Christakis, 2024), and were recruited by contacting community leaders in the region and obtaining personal informed consent. The interviews were recorded on audio in a quiet, comfortable room with a gender-balanced sample.

### Stimuli generation

We aimed to select stimuli that allowed participants to consider a specific experienced situation, because reports of one's behavior are likely less biased than widely used generalized self-reports of behavioral tendencies (Brienza et al., 2018; also see Grossmann et al., 2020; Porter et al., 2022b). The existing literature (e.g., McElroy et al., 2014) suggests that IH is especially pertinent whenever there is competition or negotiation of ideas within a relationship or group. Traditional societies are often highly hierarchical, and social norms do not allow questioning those with higher social authority (a supervisor over their employees or a husband over his wife). Therefore, the experts limited the situations of disagreement to peer relationships. Each situation prompt was followed by an event reconstruction protocol (see Brienza et al., 2018) and a set of possible reactions to the situations.

The initial item generation was conducted by English-speaking experts, with consecutive discussion of items with the native Spanish-speaking members of the team working with the Honduran population. We translated the items in Spanish using the committee approach (Harkness et al., 2004), which involved the work of experts together with context-savvy, bilingual interpreters who translated an instrument from the original language to the target language. A second bilingual translator independently back-translated the instrument from Spanish to English, and the differences from the original wordings were tracked and discussed in a team to ensure construct equivalence.

## Results

In initial cognitive interviews, we asked participants to consider different scenarios—hypothetical situations and real situations involving participants. Consistent with the principles in Table 1, Honduran participants found it easier to recall and reflect on situations from their own lives than to react to hypothetical situations. Also, though most Hondurans were able to recall meaningful events and share relevant situations, they were not comfortable when asked to discuss and reflect on disagreements with others. Respondents were also reluctant to respond when given larger numbers of IH items and disliked the perceived redundancy of certain items. Moreover, they reported difficulties with numeric responses (see OSF for the cognitive interview notes https://osf.io/qa4yf).

Based on this feedback, we updated the stimuli and repeated interviewing (see Fig. 1, also see OSF https://osf.io/6m4hc for the details of the iterative process). It led to several modifications to the initial inventory. First, we adjusted the introduction and the transitions between situations to reduce the cognitive load and facilitate understanding of our purpose and the differences across situations. Second, following feedback on the perception of repetitive and unclear questions, we shortened the inventory to six items. Third, we replaced references to "disagreement," which made some participants uncomfortable, with a more neutral "differences in opinion." Finally, we introduced a binary-chain probing procedure: For each item, interviewers first asked if participants engaged in a given action. If they did, in a follow-up probe, the interviewer queried participants about how much they engaged in each action. This way, we were able to provide an ordinal-level (not at all – a little – a lot) response for each of the six items.

## Study 2

Upon identifying suitable stimuli for our situated instrument of intellectual humility, we moved to the initial psychometric phase, where we aimed to examine the factor structure and the nomological network of the situated intellectual humility instrument. We tested an English-language version of our instrument on an online sample of US Americans, because most of the scales we relied on for convergent and discriminant validity tests were established and only validated on North American samples in English.

We expected the novel method to show its strongest convergent associations with existing trait and state IH instruments (see Table 3), with weaker links to more general or adjacent constructs such as general humility (Porter et al., 2022b). Furthermore, we expected the IH scale to be predictive of the theoretical outcomes of IH, namely greater openness and respect for other people, ideas, and experiences as indicated by higher trait openness, agreeableness, epistemic curiosity, and need for cognition and lower need for closure (e.g., McElroy et al., 2014). Since high IH is characteristic

of people who do not care too much about their intellectual dominance (Krumrei-Mancuso & Rouse, 2016; Roberts & Wood, 2007), we also expected a negative association with narcissism. Sense of mastery (self-efficacy) should be positively linked with being comfortable with recognizing and seeking feedback on one's limitations, that is, IH. Finally, a good measure of IH should be free of social desirability.

## Methods

### Data

Using Amazon Mechanical Turk (MTurk) via CloudResearch platform, we recruited 637 participants self-identified as Hispanic/Latino (49% women, 68% non-White, 25% born before 1980, 40%, and 31% in the 1980 s and 1990 s, and 2% later) who are fluent in English and residing in the USA to collect data for exploratory factor analyses.[1] Initial sample size was determined by the assumption of the moderate effect size in the multilevel designs ($r$=.20, resulting in a minimum of 100 individuals (see e.g., Kyriazos, 2018; Meuleman & Billiet, 2009) as well as by the established rules of thumb recommendation of at least 300 participants for factor analyses in individual difference research (Tabachnick et al., 2013) and further boosted with the expectation of a large dropout after data cleaning (Moss & Litman, 2019). Participants who could only recall one situation were paid $3, those who could think of two situations were paid $5, and those who completed all questions were paid $10. The data cleaning followed the open-ended questions and excluded cases with meaningless responses likely provided by bots (e.g., "Good study" when expected to report a person with a different opinion). It led to the exclusion of 280 participants. The final sample size was 357.

### Situated intellectual humility instrument

In the initial scenario, we prompted participants to reflect on the most recent instance of disagreement with a known individual. Following adjustments derived from the cognitive interviews, in the refined instrument we asked participants to recall situations involving *difference of opinion*, for which they responded to six items (see Table S1 for exact wording). Adhering to the Day/Event-reconstruction methodology (Brienza et al., 2018; Kahneman et al., 2004), participants detailed the time, place, activities, and their thoughts at the moment of the disagreement. Subsequently, they addressed our six-item questionnaire, aligning their thoughts and actions with our prompts using a two-stage binary response scale (e.g., "*Listened carefully to what the other person was saying*"*)*. Finally, to reduce ambiguity and cognitive load, we applied a branching method, converging the rating scale into a two-stage binary probe about whether participants engaged in each of the six IH-related strategies (Yes or No; if yes, "*How much did you do so*?" – *"somewhat"* versus "*very much"*). Combining responses to this binary chain allowed us to obtain an ordinal-level score for each item ("No", "Somewhat", and "Very Much").

To separate trait from state-level IH, the inventory prompted recall of three different events: general difference of opinion, a situation where they turned out to be right, and a situation where they turned out to be wrong (the latter two presented in random order). All the prompts were adapted to the unique cultural-ecological context of rural Honduras via extensive cognitive interviews and consultations with the local interpreters.

### Measures of concurrent and discriminant validity

Our survey further included ten external scales related differently to the construct of intellectual humility. To assess convergent validity, we included the General Intellectual Humility Scale (Leary et al., 2017). Despite the limitations of global self-reports highlighted in our introduction, this scale remains the most widely used benchmark in the field. Including it allows us to determine if our situated protocol captures the same core construct as established measures. Further, we administered Situated Wise Reasoning Scale (SWiS; Brienza et al., 2018), including measures of Recognition of Limits of Knowledge, Consideration of Diverse Viewpoints, Open-mindedness to Change, Search for Compromise and Conflict Resolution, and Self-transcendence/Taking the Vantage Point of an Outsider. SWiS prompted respondents to respond to the unspecified situation, the same one that was recalled in the IH inventory.

To measure discriminant validity, we also examined associations of IH to a measure of general humility (measured by Expressed Humility Scale – Landrum, 2011), traits Openness and Agreeableness (Big Five Inventory – John et al., 2008), Narcissism (NPI-16 – Ames et al., 2006), need for closure (the Brief Need for Closure Scale – Roets & Van Hiel, 2011), need for cognition (NCS-6 – Lins de Holanda Coelho et al., 2020), epistemic curiosity – including subscales of interest (joy of gaining knowledge) and deprivation (drive to close the knowledge gap – I/D Type Epistemic Curiosity scale, – Litman et al., 2010), as well as self-efficacy (measured with the General self-efficacy scale – Sherer et al., 1982). In addition, we added a measure of

[1] A technical error in Studies 2–3 prevented recording participants' specific socio-demographics. Therefore, the demographic characteristics we report are based on the initial sample estimates before exclusions provided by the survey platform.

social desirability (Final Bidimensional Impression Management Index, Blasberg et al., 2014), which contained two subdomains, Agentic and Communal Management. All measures showed a good level of internal consistency (McNeish, 2018): Reliability ω>.80 (with the exception of Consideration of diverse viewpoints subscale of the SWiS, which was.79) and all α>.75 (see SI Table S11). Following high internal consistency, we calculated composite indices for each of the criterion scales by averaging items for each given scale.

### Analytical framework

By design, the IH inventory benefited from a multilevel structure to simultaneously assess within- and between-person differences in IH, with a pragmatic consideration that exact model specifications will be detailed after the data collection and preliminary tests. We analyzed data in two complementary ways: in a two-level confirmatory factor analysis (multilevel CFA) and in a multiple-group confirmatory factor analysis (multi-group CFA).

### Two candidate measurement models of IH

We considered two candidate measurement models of IH. In Model 1, we estimated IH as a straightforward single factor where four straight-coded items were anticipated to yield positive loadings, while the two reverse-coded items were expected to load negatively, and their residuals were allowed to correlate (see Fig. 2). Model 2 offered a more nuanced approach, considering the impact of individual scale usage styles (Billiet & McClendon, 2000; Rivera-Garrido et al., 2022). Specifically, we tested a two-factor model that extended a one-factor model by adding a second uncorrelated "method" factor. Such a method factor plausibly reflects acquiescence among participants, though it might also be related to higher social desirability or satisficing (minimal-effort responding).

Given that this scale-use style could influence all items irrespective of their content, we set the factor loadings of the method factor to one for all items ("random intercept item factor model" – Maydeu-Olivares & Coffman, 2006). We considered these tests as an important step towards an unbiased IH score because the introduction of the method factor can either adjust for the response style when the latter is indeed present or bias the scores when it is not. Notably, this method factor has different meanings at different levels of analysis. As a state, it reflects higher ratings on all (even reverse-coded) items in specific situations (i.e., no differentiation between items within a situation). As a trait, it indicates an overall individual tendency to rate all six items higher (independent of situation).

### Trait and state model

Since one of the key objectives of the instrument is to capture state-level IH and separate it from the trait IH, we applied a two-level CFA (see Figure S1). First, we tested the same structure at both levels, and subsequently tested whether adding a method factor at each level separately significantly influenced model fit, and finally fixed the smallest item residuals to zero, per established practice (F. Chen et al., 2001). Following modern personality science nomenclature (Fleeson, 2001; Fleeson & Jayawickreme, 2021), we refer to the within-person factor from the multilevel CFA as a "state IH," and refer to the between-person IH factor estimates as a "trait IH." Trait IH distinguishes cross-situational independent of specific situations tendency that distinguishes between individuals. State IH, in contrast, does not distinguish between individuals but indicates responses to specific situations across all individuals. Trait IH is sensitive to individual and group differences (e.g., cultural),

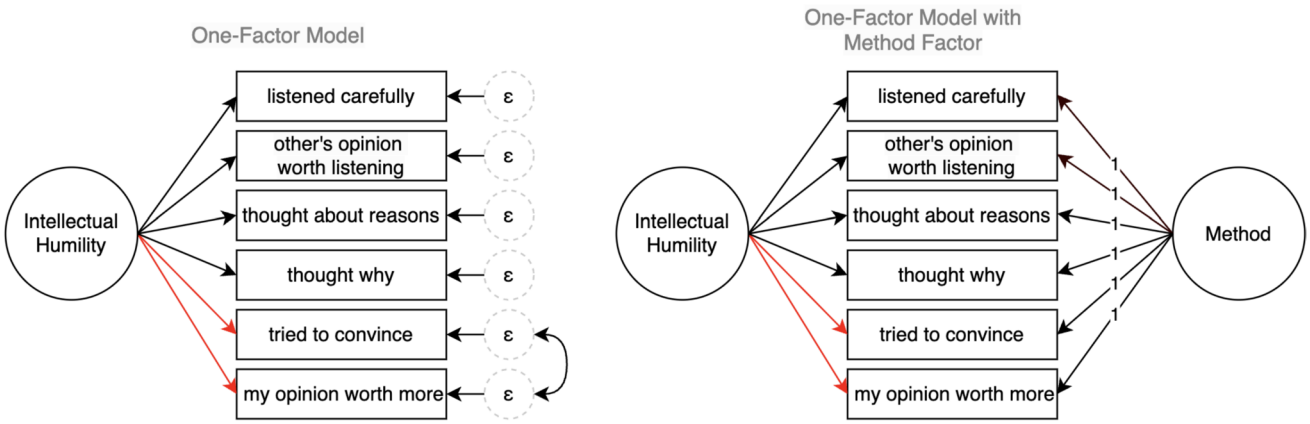


**Fig. 2** The main competing measurement models of IH

whereas state IH can point to differences across situations and ways to modulate the social environment for fostering IH (Grossmann, 2017).

The events recalled by participants in all three datasets occurred mostly within the same week (though some events also occurred within a month or were separated by up to several months, see SI Table S14). Following the typical approach of the event reconstruction method (Schwarz et al., 2009), reports on all three events were recorded during the same interview. The aggregate individual tendency across responses to these three distinct events, therefore, reflected trait IH, whereas within-individual differences across the three situations (including the unspecified one) reflected state IH. We used intraclass correlation (ICC) to indicate the share of between-person variance of IH. It was calculated as a ratio of the estimated IH factor variance at the between level to the sum of variances estimated at the between and within levels. Note that the ICC based on factor variances is already corrected for non-reliability (partialing out measurement error). In addition, we calculated ICC for each of the six items.

### Cross-sectional model

We also analyzed the aforementioned factor structures (Fig. 2) using multi-group CFA, in which groups represented the three scenarios (unspecified, "you were right," and "you were wrong"). This method enabled a formal examination of both the structural similarity (measurement invariance) and the mean differences in IH across different situations. The mean differences in IH across three situations are important criteria for the instrument's validity, showing whether IH is indeed higher in situations when one is facing limits of one's knowledge and fallibility, rather than in situations one believes to be right (and other people to be wrong).

Both multilevel and multiple group models treated indicators as ordinal variables and therefore were estimated via weighted least squares and Bayesian Gibbs sampler methods. For both approaches, we considered comparative fit index (CFI) and Tucker–Lewis Index (TLI) >.90, root mean square error of approximation (RMSEA) and standardized root mean residual (SRMR) <.08 to indicate a good enough model fit to the data (Browne & Cudeck, 1992; Hu & Bentler, 1999; Marsh et al., 2004). In the multilevel CFA framework, the SRMR was computed separately for the within- and between-individual levels.

To test for measurement invariance across situations and samples, we relied on the established recommendations in the literature (Chen, 2007; Rutkowski & Svetina, 2017), namely difference in CFI and RMSEA larger than.015 as a sign of substantial differences. When models fit the data similarly, we chose a more parsimonious model (e.g., a model in favor of invariance). Because of the large sample sizes, we did not rely on $\chi^2$ test results, as they would be invariably significant (Fan et al., 1999).

### IH scores

We obtained trait IH scores by predicting the between-individual part factor scores of IH (and, in addition, the scores of the method factor) using the multilevel CFA model described above. To do that, we applied a Bayesian estimation since the factor score prediction for the multilevel CFA was not available otherwise (see SI "Factor scores prediction in multilevel CFA" for details). To predict factor scores, we drew 1000 posterior samples of plausible values and computed a median between-individual factor score for each participant. The resulting factor score was meant to reflect trait IH. The multi-group CFA model with scalar invariance constraints predicted situation-specific scores.

## Results

Table 2 shows the model fit of different measurement models. Overall, the naïve model (without a method factor) consistently showed the worst fit, with the model fit difference being most striking for the Honduran sample. The two-factor (IH + method) model fits the data better in both multilevel CFA and multi-group CFA (Table 2 and Table S5, respectively). Moreover, in the model with the method factor, the loadings had the expected signs (i.e., were negative for reverse-coded items), whereas in a single-factor model, the loadings on the reverse-coded items were mostly non-significant. We also probed alternative factor solutions, suggesting that IH is unidimensional (see Table S2 for details).

To test whether IH construct shifts across unspecified and specified situations, we further tested for measurement invariance across the three situations (in multi-group CFA)—i.e., we examined whether the factor structure was the same across situations. The measurement invariance tests confirmed high similarity of the factor structure across situations, allowing IH mean comparison between them. The final scalar invariant model had the highest loading on "listened carefully," followed by "worth listening" and "thought about why" indicators, whereas reverse-coded items "my opinion was worth more" and "tried to convince" had the lowest magnitude of loadings (see SI Table S6 and S9).

In addition, we tested the cross-level isomorphism of the factor loadings (in multilevel CFA) to check if within-individual structure (state IH) was equivalent to the between-individual factor structure (trait IH). The results supported partial cross-level isomorphism of the two-factor model, suggesting a similar interpretation of the state and trait IH (see SI Table S10 for details).

**Table 2** Fit indices for the multilevel CFA models based on three samples

| | CFI | TLI | RMSEA | SRMR Within | SRMR Between |
|---|---|---|---|---|---|
| *English US sample* | | | | | |
| model without method factor | .957 | .919 | .038 | .049 | .041 |
| model with method factor at both levels | .961 | .927 | .036 | .049 | .035 |
| model with method factor at between-level only[a] | .961 | .930 | .035 | .049 | .044 |
| model with method factor at within-level only | .958 | .920 | .038 | .049 | .041 |
| *Spanish-language US sample* | | | | | |
| model without method factor | .987 | .975 | .025 | .038 | .070 |
| model with method factor at both levels | .984 | .970 | .027 | .036 | .079 |
| model with method factor at between-level only[b] | .982 | .971 | .027 | .038 | .091 |
| model with method factor at within-level only | .989 | .979 | .023 | .036 | .070 |
| *Honduran sample* | | | | | |
| model without method factor[a] | .912 | .844 | .056 | .059 | .078 |
| model with method factor at both levels[a] | .977 | .960 | .029 | .026 | .042 |
| model with method factor at the between level only[a] | .928 | .872 | .051 | .060 | .051 |
| model with method factor at within level only[a] | .960 | .929 | .038 | .024 | .078 |

*Note*. [a] = "worth listening" residual fixed to 0. [b] = "tried to convince" and "worth listening" residuals fixed to zero. CFI = comparative fit index, TLI = Tucker-Lewis Index, RMSEA = root mean square error of approximation, SRMR = standardized root mean residual computed separately for within- and between-person level. CFI and TLI > .90, RMSEA and SRMR < .08 indicate an acceptable model fit to the data

### Concurrent and construct validity

The correlations between IH factor scores and a set of external scales are listed in Table 3 (see also Figure S2). Consistent with our expectations for the nomological network of intellectual humility, the trait-level score of behavior-based IH showed a medium-sized association with the popular self-view measure of general IH (Leary et al., 2017), $r = .28$, $p < .001$, while the positive association with the general humility was small, $r = .12$, $p = .037$. Moreover, trait-level score of behavior-based IH showed a positive association with agreeableness and openness to experience, epistemic curiosity (interest subscale; note markedly lower effect for the deprivation subscale), and negatively with narcissism.

Our measure of IH correlated positively with the need for cognition and negatively with the need for closure. In line with expectations, we found a positive correlation of IH with general self-efficacy. Critically, and consistent with our predictions, the behavior-based IH was not associated with social desirability, whereas the method factor scores showed a positive correlation with social desirability, suggesting that the control for the method factor in our IH measurement model indeed de-biased the IH scores (for comparison of correlations, see SI Table S12).

Beyond trait-level associations, we observed a host of predicted correlations within the unspecified situation for which we also administered SWiS measure. As Table 3 shows, scores on our behavior-based IH inventory in the unspecified situation were associated with greater recognition of limits of knowledge, $r = .32$, $p < .001$, consideration of diverse viewpoints, $r = .53$, $p < .001$, open-mindedness to change, $r = .32$, $p < .001$, self-transcendence, $r = .13$, $p = .011$, as well as willingness to search for compromise and conflict resolution, $r = .32$, $p < .001$; with the exception of self-transcending these correlations were medium-large. The association of IH scores with an external measure of limits of knowledge was stronger than the correlation with self-transcendence, did not differ from the correlation with open-mindedness and the search for compromise, but the correlation with consideration of diverse perspectives was significantly stronger (see Table S12). We return to this point in the Discussion.

Notably, these correlations were limited to the common-for-both-instruments situation, whereas they did not appear when assessing IH in a different situation, highlighting a state-level sensitivity of our IH measure (see SI Table S13). Moreover, the effect sizes of situated associations were generally stronger than those of trait-level associations, suggesting the strength of situated assessment.

## Study 3

Upon identifying the best-fitting factor structure (and examining the nomological network) of our situated IH inventory among the English-speaking US residents in Study 2, we sought to perform confirmatory tests on a new sample in a different language. Specifically, we chose the Spanish-language sample—i.e., the target language of our

**Table 3** Nomological network of behavior-oriented IH in an English-speaking USA sample – Correlations with established English-language scales

| | Hypothesis | Supported | Behavior-based IH | | | Method factor | | |
|---|---|---|---|---|---|---|---|---|
| | | | *r* | 95% CI | *p* | *r* | 95% CI | *p* |
| *Trait level* | | | | | | | | |
| General IH | *Moderate positive* | + | .28 | [.18;.37] | <.001 | .07 | [–.04;.17] | .249 |
| Abstract general humility | *Weak positive* | ± | .12 | [.01;.22] | .037 | .09 | [–.01;.19] | .126 |
| BIG5 Agreeableness | *Moderate positive* | ± | .35 | [.25;.44] | <.001 | .04 | [–.07;.14] | .514 |
| BIG5 Openness | *Moderate positive* | ± | .24 | [.14;.33] | <.001 | .09 | [–.02;.19] | .157 |
| Epistemic curiosity (interest) | *Moderate positive* | ± | .27 | [.17;.36] | <.001 | .10 | [–.01;.20] | .124 |
| Epistemic curiosity (deprivation) | *Moderate positive* | – | .11 | [.00;.21] | .054 | .05 | [–.06;.15] | .410 |
| Need for cognition | *Positive* | + | .25 | [.15;.34] | <.001 | .10 | [–.01;.20] | .126 |
| Need for closure | *Negative* | + | –.22 | [–.32; –.12] | <.001 | –.002 | [–.11;.10] | .926 |
| Narcissism | *Negative* | + | –.24 | [–.34; –.14] | <.001 | .11 | [.00;.21] | .102 |
| General self-efficacy | *Weak positive* | ± | .34 | [.25;.43] | <.001 | –.07 | [–.17;.04] | .260 |
| Social desirability | *No association* | + | –.01 | [–.12;.09] | .804 | .18 | [.08;.28] | <.001 |
| *Single situation (unspecified prompt)* | | | | | | | | |
| Recognition of limits of knowledge | *Positive* | + | .32 | [.22;.41] | <.001 | .31 | [.21;.40] | <.001 |
| Consideration of diverse viewpoints | *Positive* | + | .53 | [.45;.60] | <.001 | .27 | [.18;.37] | <.001 |
| Open-mindedness to change | *Positive* | + | .29 | [.20;.39] | <.001 | .28 | [.18;.37] | <.001 |
| Search for compromise and conflict resolution | *Positive* | + | .32 | [.22;.41] | <.001 | .32 | [.22;.41] | <.001 |
| Self-transcendence | *Positive* | + | .13 | [.03;.23] | .011 | .23 | [.13;.32] | <.001 |

*Note*. IH – intellectual humility; Trait-level – between-person factor scores from the multilevel CFA; Situation-level – factor scores for the unspecified situation (the first one people recalled) from scalar multi-group CFA. Based on recent estimates of typical effects in social and personality psychology (Gignac & Szodorai, 2016), we considered $r$=.10, 20, and.30 as small, medium, and large effects, respectively

final field study. To this end, we presented a Spanish-language version of the instrument to an online self-identified Hispanic/Latino sample.

## Methods

We recruited 534 participants (51% women, 59% non-White, 21% born before 1980, 32% and 46% in the 1980s and 1990s, and 2% later) fluent in Spanish as their native language, and of Latin-American origin for confirmatory factor analysis purposes. To cast the net wide and include inexperienced MTurk participants, we did not specify pre-screening criteria (e.g., MTurk hit acceptance rate). As a result, we anticipated a substantial number of bot-like responses to exclude. After dropping 298 cases with bot-style or MTurk farmer open-ended responses (e.g., "Very good" in response to the question about the place of the event; a typical exclusion rate when no hit rates are specified; Moss & Litman, 2019), the final sample size included 236 participants—sufficient for confirmatory factor analyses in individual difference research when correlations between items are reasonably high (as established in the initial sample; Tabachnick et al., 2013).

## Results

Replicating English-speaking sample results, the model with a method factor showed better fit than the one-factor model (multi-group CFA, see SI Table S5 and S7), whereas in the multilevel CFA framework, the best fitting model included the method factor at the within-person level only (see Table 2). In fact, the two-factor model showed an even better fit than in the English-language sample. Moreover, this model supported the equivalence of structures across within- and between-individual levels – a minor deviation from the full isomorphism was due to stronger factor loadings of "thought about why" and "thought about reasons" at the between-individual level (see SI Tables S3 and S10).

Finally, the method-factor model showed scalar measurement invariance across situations (see SI Table S9). The pattern of factor loadings from the invariant model appears very similar to the pattern of loadings observed in the English online sample, except for the non-significant loading for "tried to convince" item at the between-level, suggesting its limited contribution to the overall factor on this level of analysis (see Table S7).

## Study 4

Upon establishing the structure and validity of our situation-grounded IH inventory in the English and Spanish-speaking online samples, we sought to examine the structure and validity of our approach in the target Honduran sample using face-to-face interviews. First, we tested our two measurement models against the Honduran data, then ran tests of the concurrent validity using available participant characteristics, such as socio-demographics, economic factors, and attitudes toward traditional gender and family norms, and finally explored the context of situations in relation to the associated level of IH.

### Methods

#### Data

We conducted the survey in Copán Department, Honduras, in an area of over 200 square miles of rugged mountainous terrain (Airoldi & Christakis, 2024). Most people in this area live on less than two dollars a day. A census was conducted in 176 study villages in 2015. Recruitment rates were high: government census data show approximately 32,800 eligible individuals in these villages, of whom at least 94.8% ($N = 30{,}815$) agreed to enroll in our own de novo census (in which we photographed all village residents and collected basic data about them). Of these people, 24,702 people then enrolled in a longitudinal study designed for a different purpose; the present study involved a random subsample of participants from this larger study. Specifically, we collected data in 136 systematically selected Honduran villages during 2021–2022. We solicited 20 people at random from each village. In 27 villages, fewer than 20 participants were recruited, with a minimum of 14 per village ($M = 19.0$; $SD = 2.3$). For our IH sample, the average age was 41.3 ($SD = 15.8$; range 18–93); 59% were women, 66% were partnered, and about 5% had secondary or higher education. Participants who responded "Don't know" or missed all six IH items were excluded (total of 26), resulting in the final sample size of 2567. The larger project's purposes determined the sample size, which was large enough to detect even small effects both between individuals (Tabachnick et al., 2013) and between villages (Meuleman & Billiet, 2009).

#### Measures and materials

To test for the concurrent validity of the IH scale in the Honduran sample, we investigated associations of the IH factor scores (between-person, predicted by the multilevel CFA model with method factor at both levels) with variables included in a larger study of the same participants. The primary purpose for these variables was not the testing of the IH construct validity, nevertheless some of them were proxies for the variables of interest (e.g., endorsement of violence to enforce traditional roles for women). For this reason, we expected low correlations and considered even low levels of significance. We explored gender, family norms, and demographics, including age, sex, education, and income. Measurement details of each external variable used in the concurrent validity study are listed in Table S16. To investigate associations between IH within and between villages, we used linear mixed-effects models with the IH factor score as the dependent variable and each of the variables listed above. Predictors were explicitly separated into within- and between-village parts by using within-group centering and group-level averaging, respectively. We relied on a two-level mixed-effects regression using the R package lme4 (Bates et al., 2015) to distinguish effects on IH between villages, addressing potential geographic variability in these characteristics, and within villages, addressing individual-level associations.

### Results

#### Factor model

The two-factor (IH + method factor) model again showed a substantially better fit compared to the one-factor model (see Table 2 and Table S5). Overall, the factor structure was equivalent to prior samples and supported scalar measurement invariance across situations (Table S9).

The pattern of factor loadings for the invariant model varied slightly from the models for the online samples, with the highest loadings observed for "worth listening" and "thought about reasons" items, and a negative loading of the reversed-coded item "my opinion was worth more" (see Tables S4 and S8). Importantly, when we performed formal measurement invariance tests across samples separately for each of the targeted three situations, we observed highly similar loadings and thresholds across samples (see SI for details, Table S15), confirming the scalar invariance across samples for each of the tested situations. Finally, tests of isomorphism revealed high similarity of factor loadings across within and between levels in multilevel CFA (mostly because the two reverse-coded items had stronger loadings at the between-person level on different levels; see SI Table S10).

#### Nomological network: Education, age, gender, and gender role beliefs

Associations of IH with the external variables are listed in Table 4. At the group level, intellectual humility was greater

**Table 4** Correlations and mixed effects regressions for each covariate of IH in rural Honduras

| | Correlations | | | Between-village effects | | | Within-village effects | | | *N* |
|---|---|---|---|---|---|---|---|---|---|---|
| | *r* | 95% CI | *p* | *b* | 95% CI | *p* | *b* | 95% CI | *p* | |
| *Demographics* | | | | | | | | | | |
| Age | –.002 | [–.04;.04] | .939 | .02 | [.01;.03] | .005 | .002 | [–.001;.005] | .215 | 2542 |
| Woman | –.06 | [–.10; –.02] | .001 | –.19 | [–.65;.27] | .413 | –.11 | [–.19; –.02] | .012 | 2543 |
| Education (10 steps) | .09 | [.05;.13] | <.000 | –.01 | [–.07;.05] | .764 | .05 | [.03;.07] | <.000 | 2396 |
| Income (subjective account) | .03 | [–.01;.07] | .132 | .06 | [–.17;.29] | .625 | .01 | [–.04;.07] | .636 | 2397 |
| *Gender and family norms* | | | | | | | | | | |
| Husband decision allocation (index) | –.04 | [–.08; –.002] | .041 | –.02 | [–.34;.30] | .906 | –.06 | [–.13;.01] | .086 | 2375 |
| Endorsement of violence to enforce traditional roles for women | –.06 | [–.10; –.02] | .002 | –.12 | [–.33;.10] | .284 | –.05 | [–.10; –.01] | .030 | 2343 |

*Note*. For regression analyses, each line represents a separate model with age, gender, education, and income as covariates. The dependent variable in mixed models is IH factor scores based on the model with method factor at both levels – is *z*-standardized

in villages with higher average age (supplementary results also showed that IH was greater in communities with more adults aged 60+). Furthermore, at the individual level, IH was associated with greater education (25% of Honduran participants did not have primary education, and fewer than 5% had secondary education; see Figure S4 in the SI), and with lower endorsement of violence to enforce traditional roles for women. Moreover, we observed greater IH among men than among women—a finding accounted for by the type of situation people of different genders recalled, as outlined below.

### Relational status in recalled situations and gender differences in intellectual humility

The reconstruction protocol of our IH instrument included information concerning the event specifics. The recalled situations occurred at different time points, most of them within either 1 week or 1 month prior to the measurement occasion, and almost all of them happened within 1 year (> 90%, see Table S17). Participants recalled having differences in opinion on topics of their work (19%), beliefs and religion (15%), agricultural environment (14%), mostly with their relatives (69%) and persons with the same status (65%). Based on these observations, we formally tested the association between relational status and reported intellectual humility. IH was higher among participants who tended to recall differences in opinion with persons of superior relational status (e.g., their supervisor, their parent), $\beta = .26$, $t = 5.61$, 95% CI [.17,.36], $p < .001$—i.e., having a difference of opinion with a person of senior relational status was associated with more than a quarter of standard deviation increase in IH (see Fig. 3A, B).

Notably, Honduran women were 2.93 times more likely than men to recall differences of opinion with lower status persons (e.g., their children, women = 20% vs. men = 10%, OR = 2.93, 95% CI [2.44; 3.54], $p < .001$), with whom they discussed topics concerning personal relations (15%), work (20%), health (6%) and religious matters (17%). Conversely, men more often than women recalled differences of opinion with equal status persons (e.g., friends, women = 10% vs. men = 25%; colleagues, women = 0% vs. men = 4%), and their superiors (e.g., boss, women = 0% vs. men = 3%; father, women = 2% vs. men = 5%), on topics concerning work (49%) and sports-related issues (3%). Moreover, we observed a significant indirect effect of gender on IH via average recalled status of the opponent, $b = -.07$, t = – 3.89, 95% CI [–.11, –.04], $p < .001$ (Fig. 3C). Here, the status of the opponent statistically explained the gender differences in IH—i.e., when including the relational status of the recalled partner in the same regression model, gender differences in IH became non-significant, $\beta = -.05$, $t = -1.50$, 95% CI [–.12,.02], $p = .134$. Thus, it appears that Honduran women were less likely to report IH than men because the situations they recalled involved persons of lower relational status. Thought speculatively, gender differences in the type of recalled situations are consistent with the Machismo gender norms in the region (Pérez & Morales, 2020)—with greater opportunities and license to express differences of opinion with higher status individuals among men versus women.

## Comparative analyses across Studies 2–4: The power of the situation

So far, we documented that the situated IH instrument shows robust model fit across different languages (English and Spanish) and socio-economic contexts of our samples (online crowdworkers from North America and rural farmers in Honduras). We have further demonstrated the robust nomological network of the situated measure of IH. Moreover, the context details provided as part of the event

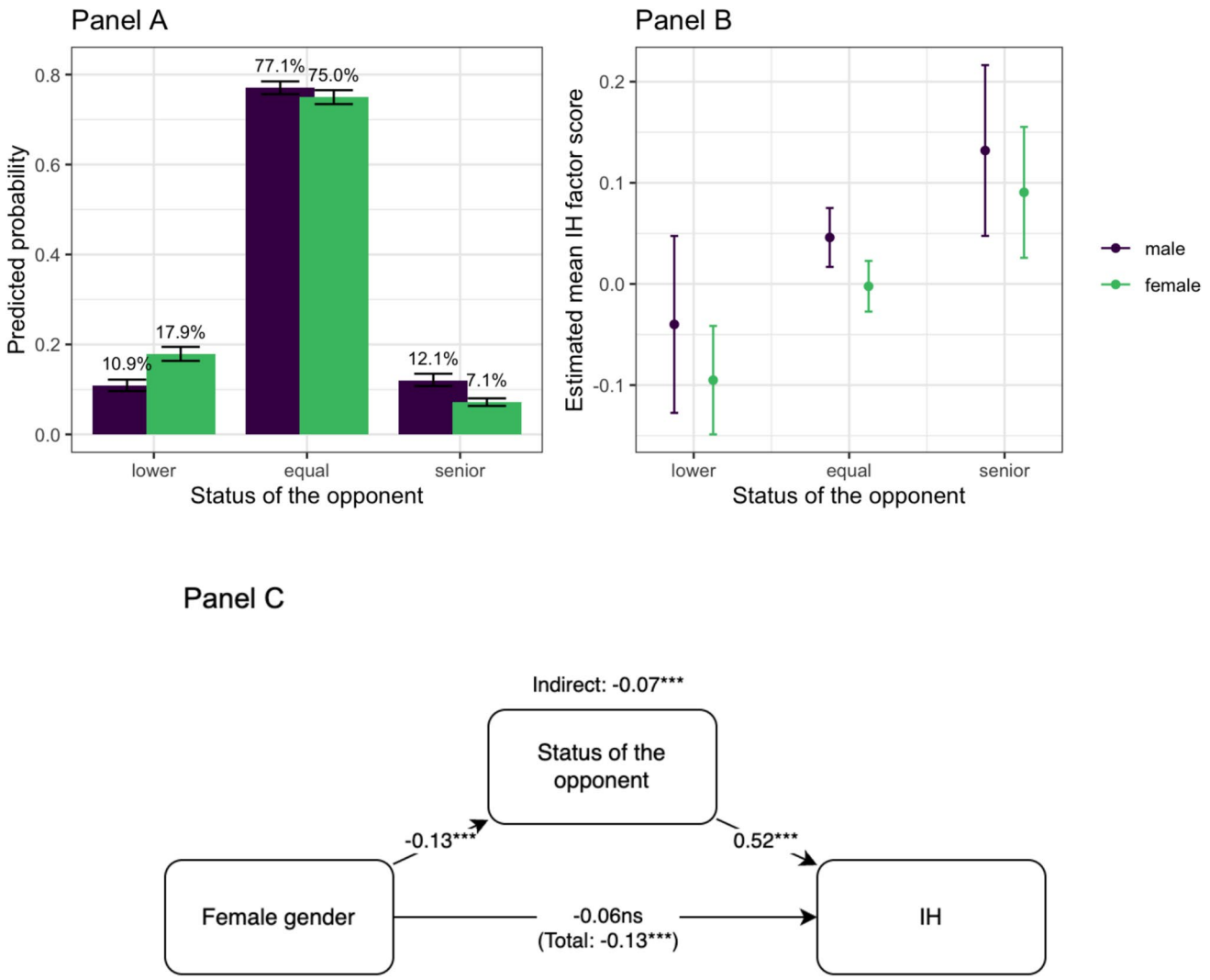


**Fig. 3** Associations between gender, status of the opponent in recalled situation, and intellectual humility (IH). Panel **A** presents the average status of the opponent across situations depending on the participant's gender. Panel **B** shows the mean IH split by participant's gender, across their opponents' status. *Vertical bars* in panels **A** and **B** represent 95% confidence interval. Panel **C** demonstrates a structural path model where control for the status of the opponent statistically accounts for the effect of gender on IH

reconstruction protocol in the situated IH instrument enabled us to probe IH differences by gender, observing systematic differences in status-related situations that men and women recalled. However, the situated nature of the present IH instrument extends beyond the study of gender differences.

Recall that we sampled responses to three distinct events. The first event was left ambiguous with no specific attribution of right or wrong. For the subsequent two events, participants were explicitly asked to recall instances where they were either right (and the other party was wrong) or vice versa, randomized to avoid order effects. This set-up afforded a unique opportunity to systematically evaluate test state-trait allocation of IH, simultaneously testing its cross-situational consistency, across different cultures. Moreover, because in each of Studies 2–4 we observed scalar invariance of the IH model across situations, we could also formally compare the average IH for each type of event. Specifically, we could compare the "default" recalled difference of opinion with the situations in which participants were explicitly instructed to recall a situation in which they were right and wrong.

## Situational variability

To explore the state-trait allocation of IH, we calculated intraclass correlation (ICC) based on sample-specific ML CFA models. ICC represents a share of variance that belongs to the between-person, or trait, level. ICC of the IH factor (adjusted for unreliability) was 26%, 19%, and 44% in the US English-, Spanish-speaking, and Honduran samples,

respectively (Fig. 4A). Thus, in the North American online samples, the share of variance in IH items belonging to the trait (between-person) was about a quarter of the total variance. In contrast, three-quarters of the variance belonged to the within-person level. Notably, the share of variance in IH belonging to the trait was higher in the Honduran sample, suggesting higher cross-situational consistency among Honduran interviewees compared to English- and Spanish-speaking online samples of US residents. It aligns with previous findings of greater consistency in responses during face-to-face interviews (Weigold et al., 2013). Furthermore, lower between-person variance in the US could result from self-selection bias in US crowdworker samples compared to a probability sample from the Honduran Copan region. Another intriguing possibility is that greater situational variance in IH may reflect broad cultural differences between more structured rural contexts in Honduras and the more flexible and diverse contexts of North America. Likewise, the IH estimated for each situation separately showed small-moderate cross-situational consistency – factor scores estimated with multi-group CFA correlated across

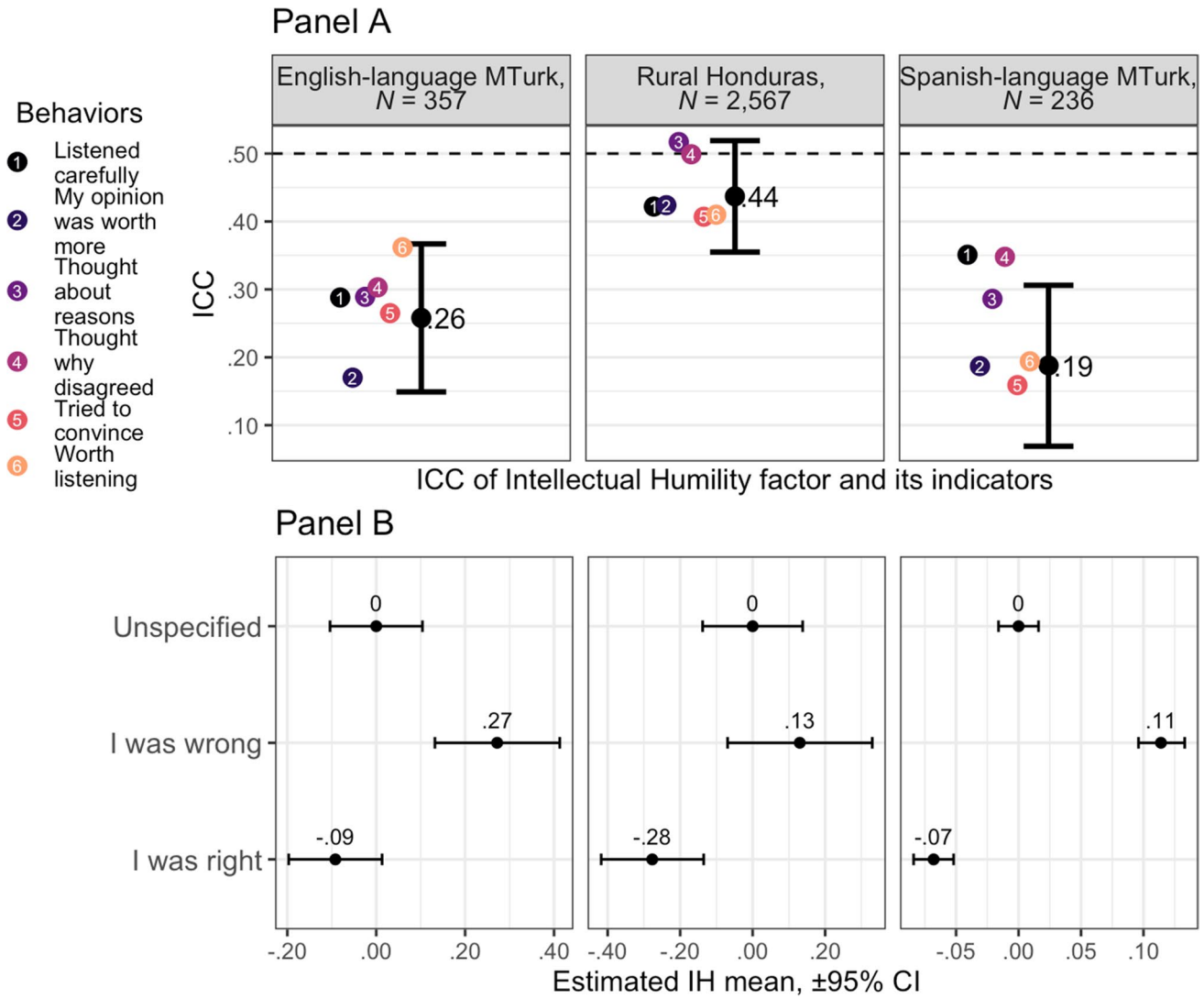


**Fig. 4** **A** Intra-class correlation (ICC) of the intellectual humility factor based on multilevel CFA, along with ICC scores for every single behavior. ICC represents the share of variance explained by between-person differences. The *dashed line* indicates 50% bound; lower ICC values in English-speaking and Honduran samples show that a minority of IH variance belongs to between-person differences. **B** Latent means of the IH factor estimated by the multi-group CFA models separately in the three samples, demonstrating criterion validity of the behavior-oriented IH method. Accounting for response bias, IH scores were highest in difference of opinion situations, one recalled being wrong as compared to situations one recalled being right. *Horizontal bars* are 95% confidence intervals (CI) estimated in the model. "Unspecified" is a reference group with its mean set to zero. CI around zero was estimated by fixing all parameters to their estimated values, but freeing factor means and variances

situations, US English:.25 ≤ *r*s <.32, Spanish:.16 < *r*s ≤.24, Honduras:.23 ≤ *r*s <.30.

### Mean differences across situations

Scores on the behavior-based IH were highest in a situation when the participant was wrong, and lowest when participant was right (Fig. 4B); the unspecified situation was falling between these two extremes. The right-wrong differences were statistically significant in each sample, US English: $b = .28$, $t = 3.40$, $p \leq .001$; Spanish: $b = .33$, $t = 2.89$, $p = .004$; Honduras: $b = .10$, $t = 9.41$, $p < .001$. These results provide strong evidence in support of the criterion validity of our measures in each culture, as one would expect higher admission of limits of knowledge and fallibility when one recalls being wrong as compared to being right.[2] Finally, we explored the variances of the IH factor across different situations (Table S14) and found that the specified prompts resulted in higher individual variability of IH, with the "You are wrong" condition having the highest variance.

## Discussion

Typical measures of metacognitive constructs, such as intellectual humility, face a paradox: those who lack intellectual humility are least likely to accurately report it, while the genuinely intellectually humble may underreport their virtues. Methodological problems compound when we try to measure intellectual humility across cultures—asking rural Honduran farmers to rate themselves on a numerical scale assumes levels of numeracy and familiarity with abstract self-evaluation that may not exist. Besides concerns about the socio-cultural representativeness of WEIRD samples used in prior IH research, global assessments invite self-enhancement, social desirability, and recall bias. These limits to dispositional methods threaten the very validity of a measurement crucial to a construct deemed important for reducing polarization, improving learning, and fostering constructive dialogue (Itzchakov et al., 2024; Porter et al., 2022b; Sgambati & Ayduk, 2023; Yang et al., 2024).

We proposed an alternative to common measurement practices: instead of asking people to self-assess how they are, we propose asking them exactly what they did in a concrete situation. Our method replaces abstract dispositional measures with concrete behavioral assessments anchored in specific, recent disagreements. Instead of asking to rate "*I accept that my beliefs and attitudes may be wrong*" on a scale from 1 to 5 (Leary et al., 2017), we suggest asking about a recent difference in opinion with someone you know and tell whether "*you listened carefully to what the other person was saying*", or not. These shifts from abstract traits to concrete behaviors and from complex numerical scales to binary-chained probes make intellectual humility a measurable metacognitive characteristic across vastly different populations.

In the present report, we provide initial evidence on the efficiency, reliability, and validity of this situated method for capturing intellectual humility. The same factor structure emerged whether administered online to English and Spanish-speaking communities in North America or face-to-face with low-literacy Honduran villagers. Our binary-chained probes yielded psychometrically sound measures showing scalar invariance across these dramatically different contexts and situations. The validity was supported by the theory-aligned person-level nomological network; intellectual humility was further associated with education, egalitarian gender norms, and higher average age village. More importantly, this method revealed something hidden by business-as-usual dispositional approaches: intellectual humility is primarily situational. In each sample, 74–81% of variance occurred within rather than between people. The same person who showed high intellectual humility when recalling being wrong turned intellectually arrogant when convinced they were right.

Notably, an unexpected finding that women in rural Honduras reported lower intellectual humility than men was fully accounted for by the relational status of the opponent: women were more likely to recall conflicts with subordinates (e.g., children), where intellectual humility tends to be lower. This observation highlights a crucial situational grounding of intellectual humility research, which we started our paper with: situations participants recall matter. In our study, Honduran men more often recalled disagreements with peers or superiors, whereas women recalled downward interactions, a pattern likely reflecting norms of gendered deference or conversational constraint, dovetailing with prior research on the inverse relationship of social status and intellectual humility (Brienza & Grossmann, 2017).

#### The humility paradox challenge

Our study started to address the humility paradox. Critically, self-assessment of humility involves not only a social

[2] We also examined the criterion validity of results without a common method factor, akin to a typical method employed in conventional scales attempting to assess IH (for a review, see Porter et al., 2022b). The results of analyses involving the IH factor without controlling for the method revealed strikingly different (and inconsistent) patterns of means across situations. For example, in Honduras, unadjusted IH did not differ between "I was wrong" and "I was right" situations and both were higher than in the unspecified situation (see Tables S18–19 and Figure S5 for details). These results suggest poor evidence of criterion validity of single-factor model without accounting for the method factor, conflating response biases with substantive IH effects.

desirability bias but also miscalibration that produces noise. High receptivity to feedback can lead some individuals to make relatively more accurate self-assessments of IH, while others with lower receptivity (that is, lower true IH) are likely to make less accurate ones. Additional sources of bias include the value that people place on IH, which may lead some individuals to underreport their actual IH. However, both underreporting and overreporting presume that people's self-evaluations meaningfully reflect their IH, whether or not they choose to disclose it. One of our central points is that some individuals may lack a self-evaluation of IH that aligns with the metacognitively oriented understanding of the construct and is not confounded with IH (Porter et al., 2022b).

Our proposed measurement approach was designed to accommodate this possibility, as well as the various biases and sources of noise that may arise. It continues the line of the existing attempts to situationally ground intellectual humility measurement (Hoyle et al., 2016; Zachry et al., 2018), applying a principled and generalizable approach (see Table 1).

Across three studies, we showed that intellectual humility is a measurable and meaningful psychological construct across vastly different cultural and educational settings—*if* it is assessed via concrete behaviors, in ways that reflect people's lived social experiences. By advancing a context-sensitive, situational approach to intellectual humility, we offer a methodologically robust and philosophically grounded framework for future research on how people manage their knowledge limits in everyday life. These insights open the door to broader inquiries into intellectual humility's role in social cohesion, learning, conflict resolution, and collective adaptation worldwide.

### Limitations and future directions

Several caveats are in order before concluding. Our samples were limited to online US and in-person rural Honduran participants. Although we observed consistent results across these vastly different samples, it is yet unknown how these findings would generalize beyond them. Additionally, although the situated behavior approach helped alleviating effects of self-presentation, it still relied on self-reports of one's behavior, albeit such reports were specific and concerned concrete actions. Here, one particular item, "tried to convince," stood out and did not receive significant factor loadings in some of our measurement models. This item involves motivations and verbal behavior, whereas the other five more squarely tap into internal metacognitive processes.[3] Future research could explore how adding more purely behavioral items could form a distinct dimension of IH and how performance on this dimension interacts with the metacognitive IH core across different social, political, and epistemic domains of one's life. Future work can also explore convergent evidence from partner- and observer-provided ratings, as well as non-verbal cues (e.g., body language, eye movements).

Finally, our instrument relied on targeted event reconstruction rather than random experience sampling. While prompts such as "when you were wrong" effectively acted as sampling quotas to mitigate self-selection bias, this design limited the scope of our findings in two ways. First, our protocol was strictly interpersonal, focusing on the domain of difference of opinion with another person. As noted in recent frameworks, IH contains both an internal metacognitive core and expressed social components (Porter et al., 2022a, b). Because our measure targeted how individuals navigate epistemic friction with others, it naturally tracked socially expressed facets (e.g., SWiS *Consideration of Diverse Viewpoints*, $r = .53$) rather than private metacognition (SWiS *Limits of Knowledge*, $r = .32$). Second, while we captured variance across outcomes (being right vs. wrong), we did not systematically vary other critical dimensions such as perceived threat, epistemic uncertainty, or social stakes (Grossmann & Kross, 2014; Porter et al., 2022b). Consequently, the current instrument likely over-represents the expression of IH in resolved situations and under-represents it in continuing uncertain circumstances. To map the full 'density profile' of the trait (Fleeson & Jayawickreme, 2021), future research should extend this situated protocol beyond interpersonal disagreements to include solitary tasks (to isolate internal metacognition) and situations characterized by high threat or ambiguity (to stress-test the metacognitive core).

## Conclusion

We argue that any assessment of metacognitive constructs can benefit from the principles applied in the present research: tying responses to the concrete situations, capturing specific behaviors in construct- and culture-relevant contexts with theory-aligned prompts to test for the situational variability. By expanding the focus of measurement beyond abstract self-evaluations and embracing contextually anchored assessments, psychological science can more accurately map the metacognitive landscape and inform interventions aimed at fostering open-mindedness, reducing polarization, and promoting constructive dialogue across groups.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.3758/s13428-026-03016-x.

[3] Higher factor loadings from the metacognitive items (compared to the non-significant loading of the behavior-oriented "tried to convince" item) on the overall construct empirically support the theoretical position that internal metacognition rather than expressed behavior is at the center of intellectual humility (Porter et al., 2022a, b).

**Authors' contributions** NC, RB, and IG conceived and designed the experiments. RB and NC performed the experiments. MR and ALRR analyzed the data. MR and IG contributed materials and analysis tools. MR, NC, ALRR, and IG wrote the paper.

**Funding** The study was supported by The John Templeton Foundation #61921, Bill and Melinda Gates Foundation #OPP1098684 and Nomis Foundation (to NC), The John Templeton Foundation #62260, Templeton World Charity Foundation #TWCF-2023–32568, Social Sciences and Humanities Research Council of Canada Insight grant #435-2014-0685 (to IG).

**Data availability** Data for the online samples is publicly available at the OSF directory (https://osf.io/wj3fe). The data file for the in-person Honduran interviews is published partially to preserve the privacy of the participants, as it contains identifying information including age, gender, and education but the published data is sufficient for the replication of the factor structure.

**Code availability** The replication materials, including study plan, stimuli, qualitative summaries, and model reproduction codes, can be found in the OSF directory: https://osf.io/wj3fe.

## Declarations

**Ethics approval** Studies 2 and 3 were approved by the Yale Human Subjects Committee, IRB protocol 2000030316 on April 19, 2021. Study 4 was approved by the Yale Human Subjects Committee, IRB protocol 2000022813, April 13, 2018.

**Consent to participate** Participants provided informed consent before participating in the study.

**Consent for publication** Not applicable.

**Conflicts of interest/Competing interests** Not applicable.

## References

Airoldi, E. M., & Christakis, N. A. (2024). Induction of social contagion for diverse outcomes in structured experiments in isolated villages. *Science, 384*(6695), eadi5147.

Alfano, M., Iurino, K., Stey, P., Robinson, B., Christen, M., Yu, F., & Lapsley, D. (2017). Development and validation of a multidimensional measure of intellectual humility. *PLoS ONE, 12*(8), e0182950.

Ames, D. R., Rose, P., & Anderson, C. P. (2006). The NPI-16 as a short measure of narcissism. *Journal of Research in Personality, 40*(4), 440–450.

Ballantyne, N. (2023). Recent work on intellectual humility: A philosopher's perspective. *The Journal of Positive Psychology, 18*(2), 200–220.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48.

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly, 71*(2), 287–311.

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling, 7*(4), 608–628.

Blasberg, S. A., Rogers, K. H., & Paulhus, D. L. (2014). The Bidimensional Impression Management Index (BIMI): Measuring agentic and communal forms of impression management. *Journal of Personality Assessment, 96*(5), 523–531.

Bostyn, D. H., Sevenhant, S., & Roets, A. (2018). Of mice, men, and trolleys: Hypothetical judgment versus real-life behavior in trolley-style moral dilemmas. *Psychological Science, 29*(7), 1084–1093.

Brienza, J. P., & Grossmann, I. (2017). Social class and wise reasoning about interpersonal conflicts across regions, persons and situations. *Proceedings of the Royal Society b: Biological Sciences, 284*(1869), 20171870.

Brienza, J. P., Kung, F. Y., Santos, H. C., Bobocel, D. R., & Grossmann, I. (2018). Wisdom, bias, and balance: Toward a process-sensitive measurement of wisdom-related cognition. *Journal of Personality and Social Psychology, 115*(6), 1093.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21*(2), 230–258.

Chen, F. F. (2007). Sensitivity of goodness-of-fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504.

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods & Research, 29*(4), 468–508.

Costello, T. H., Newton, C., Lin, H., & Pennycook, G. (2023). *Intellectual humility questionnaires fail to predict metacognitive skill: Implications for theory and measurement*.

Dunning, D. (2011). The Dunning–Kruger effect: On being ignorant of one's own ignorance. In *Advances in experimental social psychology* (Vol. 44, pp. 247–296). Elsevier.

Dorfman, A., Moscovitch, D. A., Chopik, W. J., & Grossmann, I. (2022). None the wiser: Year-long longitudinal study on effects of adversity on wisdom. *European Journal of Personality, 36*(4), 559–575.

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 56–83.

Fiske, A. P. (1992). The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review, 99*(4), 689.

Fleeson, W. (2001). Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology, 80*(6), 1011.

Fleeson, W., & Jayawickreme, E. (2021). Whole traits: Revealing the social-cognitive mechanisms constituting personality's central variable. In *Advances in experimental social psychology* (Vol. 63, pp. 69–128). Elsevier.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience, 8*, 443.

Funder, D. C. (2006). Towards a resolution of the personality triad: Persons, situations, and behaviors. *Journal of Research in Personality, 40*(1), 21–34.

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74–78.

Grossmann, I. (2017). Wisdom in context. *Perspectives on Psychological Science, 12*(2), 233–257.

Grossmann, I., & Kross, E. (2014). Exploring Solomon's paradox: Self-distancing eliminates the self-other asymmetry in wise reasoning about close relationships in younger and older adults. *Psychological Science, 25*(8), 1571–1580.

Grossmann, I., & Na, J. (2014). Research in culture and psychology: Past lessons and future challenges. *Wiley Interdisciplinary Reviews: Cognitive Science, 5*(1), 1–14.

Grossmann, I., Weststrate, N. M., Ardelt, M., Brienza, J. P., Dong, M., Ferrari, M., Fournier, M. A., Hu, C. S., Nusbaum, H. C., & Vervaeke, J. (2020). The science of wisdom in a polarized world: Knowns and unknowns. *Psychological Inquiry, 31*(2), 103–133.

Grossmann, I., Smith, J. A., Wegenschimmel, N., Diep, P., & Dorfman, A. (2025). Wisdom in time: Advancing dynamic models of wisdom, intellectual humility, open-mindedness, and perspective-taking. *European Review of Social Psychology*, 1–36.

Harkness, J., Pennell, B.-E., & Schoua-Glusberg, A. (2004). Survey questionnaire translation and assessment. *Methods for Testing and Evaluating Survey Questionnaires*, 453–473.

Hazlett, A. (2012). Higher-order epistemic attitudes and intellectual humility. *Episteme, 9*(3), 205–223.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–83.

Hilbert, S., Küchenhoff, H., Sarubin, N., Toyo Nakagawa, T., & Bühner, M. (2016). The influence of the response format in a personality questionnaire: An analysis of a dichotomous, a Likert-type, and a visual analogue scale. *TPM: Testing, Psychometrics, Methodology in Applied Psychology*, *23*(1).

Hoyle, R. H., Davisson, E. K., Diebels, K. J., & Leary, M. R. (2016). Holding specific views with humility: Conceptualization and measurement of specific intellectual humility. *Personality and Individual Differences, 97*, 165–172.

Hruschka, D. J., Munira, S., Jesmin, K., Hackman, J., & Tiokhin, L. (2018). Learning from failures of protocol in cross-cultural research. *Proceedings of the National Academy of Sciences, 115*(45), 11428–11434.

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

Itzchakov, G., Reis, H. T., & Rios, K. (2024). Perceiving others as responsive lessens prejudice: The mediating roles of intellectual humility and attitude ambivalence. *Journal of Experimental Social Psychology, 110*, 104554.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy. *Handbook of Personality: Theory and Research, 3*(2), 114–158.

Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A survey method for characterizing daily life experience: The day reconstruction method. *Science, 306*(5702), 1776–1780.

Kleitman, S., Hui, J.S.-W., & Jiang, Y. (2019). Confidence to spare: Individual differences in cognitive and metacognitive arrogance and competence. *Metacognition and Learning, 14*(3), 479–508.

Koetke, J., Schumann, K., & Porter, T. (2022). Intellectual humility predicts scrutiny of COVID-19 misinformation. *Social Psychological and Personality Science, 13*(1), 277–284. https://doi.org/10.1177/1948550620988242

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*(3), 213–236.

Krumrei-Mancuso, E. J., & Rouse, S. V. (2016). The development and validation of the comprehensive intellectual humility scale. *Journal of Personality Assessment, 98*(2), 209–221.

Krumrei-Mancuso, E. J., Haggard, M. C., LaBouff, J. P., & Rowatt, W. C. (2020). Links between intellectual humility and acquiring knowledge. *The Journal of Positive Psychology, 15*(2), 155–170. https://doi.org/10.1080/17439760.2019.1579359

Krumrei-Mancuso, E. J., Pärnamets, P., Bland, S., Astola, M., Cichocka, A., de Ridr, J., Mercier, H., Meyer, M., O'Connor, C., Porter, T., de Ridr, J., Tanesini, A., Alfano, M., & Van Bavel, J. J. (2025). Toward an understanding of collective intellectual humility. *Trends in Cognitive Sciences, 29*(1), 15–27.

Kube, T., & Rozenkrantz, L. (2021). When beliefs face reality: An integrative review of belief updating in mental health and illness. *Perspectives on Psychological Science, 16*(2), 247–274.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480.

Kyriazos, T. A. (2018). Applied psychometrics: Sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. *Psychology, 9*(08), 2207.

Landrum, R. E. (2011). Measuring dispositional humility: A first approximation. *Psychological Reports, 108*(1), 217–228.

Leary, M. R., Diebels, K. J., Davisson, E. K., Jongman-Sereno, K. P., Isherwood, J. C., Raimi, K. T., Deffler, S. A., & Hoyle, R. H. (2017). Cognitive and interpersonal features of intellectual humility. *Personality and Social Psychology Bulletin, 43*(6), 793–813.

Leman, J., Kurinec, C., & Rowatt, W. (2023). Overconfident and unaware: Intellectual humility and the calibration of metacognition. *The Journal of Positive Psychology, 18*(1), 178–196.

Lenzner, T., Neuert, C., & Otto, W. (2016). *Cognitive Pretesting (Version 2.0)*.

Lewin, K. (1936). A dynamic theory of personality: Selected papers. *The Journal of Nervous and Mental Disease, 84*(5), 612–613.

Lins de Holanda Coelho, G., HP Hanel, P., & J. Wolf, L. (2020). The very efficient assessment of need for cognition: Developing a six-item version. *Assessment, 27*(8), 1870–1885.

Litman, J. A., Crowson, H. M., & Kolinski, K. (2010). Validity of the interest-and deprivation-type epistemic curiosity distinction in non-students. *Personality and Individual Differences, 49*(5), 531–536.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review, 98*(2), 224.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*(3), 320–341.

Masuda, T., & Nisbett, R. E. (2001). Attending holistically versus analytically: Comparing the context sensitivity of Japanese and Americans. *Journal of Personality and Social Psychology, 81*(5), 922.

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods, 11*(4), 344.

McElroy, S. E., Rice, K. G., Davis, D. E., Hook, J. N., Hill, P. C., Worthington, E. L., Jr., & Van Tongeren, D. R. (2014). Intellectual humility: Scale development and theoretical elaborations in the context of religious leadership. *Journal of Psychology and Theology, 42*(1), 19–30.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412.

Metcalfe, J., & Finn, B. (2013). Metacognition and control of study choice in children. *Metacognition and Learning, 8*, 19–46.

Meuleman, B., & Billiet, J. (2009). A Monte Carlo sample size study: How many countries are needed for accurate multilevel SEM? *Survey Research Methods, 3*(1), 45–58.

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102*(2), 246.

Molenaar, P. C. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement, 2*(4), 201–218.

Moss, A., & Litman, L. (2019). After the Bot Scare: Understanding What's Been Happening With Data Collection on MTurk and How to Stop It. *CloudResearch Blog*. https://www.cloudresearch.com/resources/blog/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-to-stop-it/

Murray, M. H., & Blume, J. D. (2021). FDRestimation: Flexible false discovery rate computation in R. *F1000Research*. https://doi.org/10.12688/f1000research.52999.2

Muthén, L. K., & Muthén, B. O. (2022). *Mplus: Statistical analysis with latent variables; user's guide;[version 8]*. Muthén et Muthén.

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review, 108*(2), 291.

Paulhus, D. L., & Reid, D. B. (1991). Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology, 60*(2), 307.

Pérez, O. F. R., & Morales, A. (2020). Machismo. *The Wiley Encyclopedia of Personality and Individual Differences: Clinical, Applied, and Cross-Cultural Research*, 243–246.

Poortinga, Y. H., & Van de Vijver, F. J. (2004). Cultures and Cognition: Performance Differences and Invariant Structures. In R. Sternberg & E. L. Grigorenko (Eds.), *Culture and competence: Contexts of life success* (pp. 139–162). American Psychological Association.

Porter, T., Baldwin, C. R., Warren, M. T., Murray, E. D., Cotton Bronk, K., Forgeard, M. J., ... & Jayawickreme, E. (2022a). Clarifying the content of intellectual humility: A systematic review and integrative framework. *Journal of Personality Assessment, 104*(5), 573–585.

Porter, T., Elnakouri, A., Meyers, E. A., Shibayama, T., Jayawickreme, E., & Grossmann, I. (2022b). Predictors and consequences of intellectual humility. *Nature Reviews Psychology, 1*(9), 524–536. https://doi.org/10.1038/s44159-022-00081-9

R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rivera-Garrido, N., Ramos-Sosa, MdelP., Accerenzi, M., & Brañas-Garza, P. (2022). Continuous and binary sets of responses differ in the field. *Scientific Reports, 12*(1), 14376.

Roberts, R. C., & Wood, W. J. (2007). *Intellectual virtues: An essay in regulative epistemology*. OUP UK.

Roets, A., & Van Hiel, A. (2011). Item selection and validation of a brief, 15-item version of the Need for Closure Scale. *Personality and Individual Differences, 50*(1), 90–94.

Rutkowski, L., & Svetina, D. (2017). Measurement invariance in international surveys: Categorical indicators and fit measure performance. *Applied Measurement in Education, 30*(1), 39–51.

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology, 45*(3), 513.

Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior: Cognition, communication, and questionnaire construction. *The American Journal of Evaluation, 22*(2), 127–160.

Schwarz, N., Kahneman, D., Xu, J., Belli, R., Stafford, F., & Alwin, D. (2009). Global and episodic reports of hedonic experience. *Using Calendar and Diary Methods in Life Events Research*, 157–174.

Sedikides, C., & Gregg, A. P. (2008). Self-enhancement: Food for thought. *Perspectives on Psychological Science, 3*(2), 102–116.

Sgambati, T. J., & Ayduk, O. N. (2023). Intellectual humility and political polarization: An exploration of social networks, attitudes, and affect. *Political Psychology, 44*(4), 807–828.

Sherer, M., Maddux, J. E., Mercandante, B., Prentice-Dunn, S., Jacobs, B., & Rogers, R. W. (1982). The self-efficacy scale: Construction and validation. *Psychological Reports, 51*(2), 663–671.

Smith, J. A., Dorfman, A., Wegenschimmel, N., & Grossmann, I. (2025). Wisdom reconsidered: A dynamic network account of metacognition and complex thought. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001821

Staudinger, U. M., & Baltes, P. B. (1996). Interactive minds: A facilitative setting for wisdom-related performance? *Journal of Personality and Social Psychology, 71*(4), 746.

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2013). *Using multivariate statistics* (Vol. 6). Pearson Boston, MA.

Varnum, M. E., Grossmann, I., Kitayama, S., & Nisbett, R. E. (2010). The origin of cultural differences in cognition: The social orientation hypothesis. *Current Directions in Psychological Science, 19*(1), 9–13.

Weigold, A., Weigold, I. K., & Russell, E. J. (2013). Examination of the equivalence of self-report survey-based paper-and-pencil and internet data collection methods. *Psychological methods, 18*(1), 53.

Whitcomb, D., Battaly, H., Baehr, J., & Howard-Snyder, D. (2017). Intellectual humility: Owning our limitations. *Philosophy and Phenomenological Research, 94*(3), 509–539. https://doi.org/10.1111/phpr.12228

Yang, Z. R., Kung, F. Y., Brienza, J. P., & Chao, M. M. (2024). Bridging social divides: The role of wise reasoning in improving intergroup relations. *Translational Issues in Psychological Science, 10*(1), 69.

Zachry, C. E., Phan, L. V., Blackie, L. E., & Jayawickreme, E. (2018). Situation-based contingencies underlying wisdom-content manifestations: Examining intellectual humility in daily life. *The Journals of Gerontology, Series b: Psychological Sciences and Social Sciences, 73*(8), 1404–1415.