Supplementary Material A. Naive Judge Content Validation: Prolific Implementation Protocol

Sections 1–7 reproduce the participant-facing text presented to naive judges in Qualtrics. Any text explicitly labelled 'Researcher-only (not shown to participants)' documents recruitment, programming, or analysis decisions and was not displayed to participants. Section 8 contains the analysis framework and decision rules and was not shown to participants.

Section 1. Information Sheet and Consent

[SCREEN 1]

Box A1. Screen 1: Information sheet and consent text.

Study: Categorizing Investor Experiences with AI Tools

Purpose. You will be shown short statements from a draft survey about how traders experience using AI tools (large language models, 'LLMs') when making trading decisions. Your task is to act as a judge: you will read brief definitions of several psychological constructs and indicate which construct each statement best represents.

What you will do. You will: (1) answer screening questions about age and trading experience; (2) read definitions of five constructs; (3) answer brief comprehension questions; and (4) for 22 items (20 content statements + 2 attention/data-quality checks), indicate which construct each best matches and rate the fit.

Important. This is not financial advice. Please do not enter any personally identifying or sensitive financial information (for example, account numbers).

Time and payment. Planned: approximately 12–15 minutes. [1] Payment follows Prolific standard rates.

Voluntary and confidential. Participation is voluntary and you may stop at any time. Responses are linked only to an anonymous Prolific ID and will be used for academic research.

Contact. Questions: [PI email]

Consent Question

Do you agree to take part in this academic study under these terms?

○ Yes, I agree [Continue]
○ No, I do not agree [End survey]

Section 2. Eligibility Screener

[SCREEN 2]

Prolific Pre-screen Settings (Researcher-only, not shown to participants): Approval rate ≥ 95%; participants resident in OECD[2] countries.

S0. English Fluency

Are you fluent in English?
○ Yes [Continue]
○ No [End survey]

[1] Actual: approximately 20–25 minutes (median observed duration in the final sample was 22.2 minutes).
[2] OECD (Organisation for Economic Co-operation and Development), https://www.oecd.org/en.html

S1. Age

How old are you (in years)? [Open numeric entry; must be ≥ 18] [3]

S2. Trading Frequency

In the past 12 months, how often have you traded stocks, ETFs, options, or crypto-assets using real money?

○ Daily or almost daily [Continue]
○ Weekly [Continue]
○ Monthly [Continue]
○ A few times per year, e.g., above 10 trades for the past 12 months [Continue]
○ Rarely, e.g., below 10 trades for the past 12 months [End survey]
○ Never [End survey]

S3. Trading Role

Which best describes you?

○ I trade only my own money (retail trader) [Continue]
○ I trade as part of my job for a financial firm or clients only [End survey]
○ I do not trade with real money or only use demo accounts [End survey]

S4. LLM Usage for Trading

Have you ever used a Generative AI / LLM text tool(s) (like ChatGPT, Claude, Gemini, or Perplexity) to ask a finance, market, or trading-related question?

○ Yes [Continue]
○ No [End survey]

S5. Occupation

What is your main occupation? (e.g., 'software engineer', 'university student', 'retired')
[Open text]

Researcher-only (not shown to participants): Duplicate submissions (analysis-stage rule): If multiple Qualtrics submissions share the same PROLIFIC_PID, only the earliest submission is retained ("first-response rule"). Later submissions are excluded, especially where eligibility-screen responses change in a direction that enables access to the task (e.g., changing LLM-use screening from "No" to "Yes"). This rule is applied to prevent screening evasion and aligns with standard handling of duplicate panel submissions.

Section 3. Construct Definitions[4]

[SCREEN 3 – Definitions permanently visible in sidebar/sticky header throughout task]

Instructions: Please read these definitions carefully. You will need to use them to categorize the statements on the next screens. These definitions will remain visible at the top of every screen during the task.

Researcher-only (not shown to participants): Implementation notes: (1) The order of construct definitions should be randomised across participants to control for primacy effects. (2) Definitions must be displayed in a permanent sidebar or sticky header on every item screen; do not rely on a 'Show Definitions' button. (3) Letter labels (A, B, C…) are omitted to allow independent randomisation of definition order and response option order.

Perceived Cognitive Assistance (PCA)

[3] Age was enforced via Prolific profile filters and not asked in Qualtrics
[4] Researcher-only (not shown to participants)

Box A2. Screen 3: Construct definition text shown to participants (Perceived Cognitive Assistance, PCA).

PCA is the trader's perception that using an LLM acts as a cognitive scaffold—structuring the reasoning process and expanding mental processing capacity when making trading decisions.

It is about feeling that the LLM:

- Provides a structured path from an initial idea to an executable order
- Allows you to process more information without cognitive overload
- Supports your active scrutiny of your own plans (e.g., noticing gaps or inconsistencies)
- Helps you simulate and compare scenarios
- Facilitates reflection on your own decision process

Key point: PCA concerns the *mechanism of thinking*—the scaffolding, structure, and filtering that supports *your own reasoning* at the moment of decision. It is *not* about outcomes (that's PU), interface effort (that's PEOU), or system reliability (that's Trust).

Key indicator: Statements focus on your thinking process and your active role (e.g., 'The tool helps *me* structure...', 'I become more effective at spotting...').

Perceived Usefulness (PU)

Box A3. Screen 3: Construct definition text shown to participants (Perceived Usefulness, PU).

PU is the trader's belief that using an LLM leads to advantageous trading outcomes, financial performance, and efficiency gains.

A trader high in PU thinks that:

- Using the LLM improves their trading results
- The tool is beneficial and worth using for trading tasks
- It saves time, increases productivity, or boosts performance

Key point: PU concerns beliefs about *outcomes and results after* using the tool—profits, returns, time saved, performance improvements. Statements emphasizing *how you think during* decision-making are PCA, not PU.

Key indicator: Statements focus on results and value (e.g., 'The tool saves me time', 'The tool improves my returns', 'The tool is beneficial', 'The tool is profitable').

Perceived Ease of Use (PEOU)

Box A4. Screen 3: Construct definition text shown to participants (Perceived Ease of Use, PEOU).

PEOU is the trader's belief that using an LLM for trading is easy and low-effort.

It is about:

- How easy it is to learn the tool
- How clear and understandable the interface and responses are
- Getting the tool to do what you want without frustration

Key point: PEOU concerns the effort of interacting with the *software itself* (learning, prompting, and understanding responses). It does *not* concern whether the trading decision feels cognitively clearer (that would be PCA).

Important distinction: Words like 'clearer' or 'easier' in the context of *understanding a trading situation* (the cognitive task) are about PCA, not PEOU. PEOU is only about how easy the *software interaction* is.

Key indicator: Statements focus on the interface and software interaction (e.g., 'The tool is easy to learn', 'The prompts are straightforward', 'The responses are readable').

Trust in the LLM

Box A5. Screen 3: Construct definition text shown to participants (Trust in the LLM).

Trust is the trader's confidence that the LLM's internal operations are accurate and that it functions correctly without requiring constant user supervision.

A trader high in Trust believes:

- The tool's outputs are generally dependable and accurate
- It will not behave in surprising or unsafe ways
- It is comfortable to rely on the tool as a starting point

Key point: Trust concerns confidence in the *system's reliability and accuracy*. It is *passive* confidence that the system is correct. Statements describing the *user's active* checking, vigilance, or improved error detection are PCA, not Trust.

Critical distinction from PCA: Trust is about whether *the tool itself* is accurate and reliable (system state). PCA is about whether the tool helps *you* become more effective at checking your own work (user agency). 'The tool catches errors' = Trust. 'The tool helps *me* catch errors' = PCA.

Key indicator: Statements focus on the system's reliability (e.g., 'The tool is accurate', 'I can rely on its outputs', 'The tool behaves predictably', 'The tool is safe').

Trading Self-Efficacy (TSE)

Box A6. Screen 3: Construct definition text shown to participants (Trading Self-Efficacy, TSE).

TSE is the trader's confidence in their own baseline ability to plan, execute, and manage trades—reflecting self-perceived competence rather than assistance provided by any tool.

A trader high in TSE feels capable of:

- Organising the steps of a trade
- Coping with pressure and uncertainty
- Recognising when more information is needed
- Learning from past trades and adjusting

Key point: TSE concerns the trader's belief in their *own skills and judgment*—what they bring to trading as a person. It is about *who I am as a trader*, not about what a tool does for me.

Key indicator: Statements emphasise the trader's inherent capability, confidence, or skill (e.g., 'I am good at…', 'I can handle…', 'I know how to…'). Statements that emphasise what a tool *provides or enables* point toward other constructs.

Section 4. Comprehension Check

[SCREEN 4]

Instructions: Before we begin, please answer these questions to confirm you understand the differences between the categories.

C1. Interface vs. Cognitive Clarity

Which construct is most about how easy it feels to *learn and operate the LLM software*, rather than how much it improves your thinking or outcomes?

- ○ Perceived Cognitive Assistance (PCA)
- ○ Perceived Usefulness (PU)
- ○ Perceived Ease of Use (PEOU) **[CORRECT]**[5]
- ○ Trust in the LLM

[5] **[CORRECT] -** Researcher-only (not shown to participants)

○ Trading Self-Efficacy (TSE)

C2. Own Skills vs. Tool-Assisted

Which construct is most about the trader's *own trading skills and confidence*, regardless of whether they use an LLM?

○ Perceived Cognitive Assistance (PCA)
○ Perceived Usefulness (PU)
○ Perceived Ease of Use (PEOU)
○ Trust in the LLM
○ Trading Self-Efficacy (TSE) **[CORRECT]**

C3. Enhanced Thinking Process

Which construct focuses on feeling that the LLM *helps you think through* a complex problem more clearly?

○ Perceived Cognitive Assistance (PCA) **[CORRECT]**
○ Perceived Usefulness (PU)
○ Perceived Ease of Use (PEOU)
○ Trust in the LLM
○ Trading Self-Efficacy (TSE)

C4. Tool Catches Errors vs. You Catch Errors

Consider these two statements: (A) 'The LLM catches errors in my trading plans.' (B) 'The LLM helps me catch errors in my trading plans.' Which answer best describes the difference?

○ Both are Trust statements
○ Both are PCA statements
○ Statement A is Trust; Statement B is PCA **[CORRECT]**
○ Statement A is PCA; Statement B is Trust
○ Both are Perceived Usefulness statements

C5. Thinking Process vs. Trading Results

Which construct is primarily about whether using the LLM leads to *better financial results and improved trading outcomes*?

○ Perceived Cognitive Assistance (PCA)
○ Perceived Usefulness (PU) **[CORRECT]**
○ Perceived Ease of Use (PEOU)
○ Trust in the LLM
○ Trading Self-Efficacy (TSE)

Researcher-only (not shown to participants): Routing and comprehension checks. The comprehension items were implemented as a forced-correct instructional gate. Participants could not proceed until selecting the correct option. If an incorrect option was selected, the survey displayed corrective feedback and required another attempt before advancing. In the Qualtrics export used for analysis, only the final (correct) response was retained for these items. Therefore, comprehension-check performance was not analysed, and no exclusions or robustness checks based on comprehension checks were applied.

Section 5. Main Classification Task

[One screen per statement, randomised order; 22 classification screens total = 20 content items (16 PCA candidates + 4 fillers) + 2 attention-check screens.]

Instructions: You will now see 22 statements (20 content statements + 2 attention/data-quality checks), presented one at a time in randomised order. For each one: (1) Choose which construct it best

represents. (2) Rate how well it fits. (3) Rate overlap with other constructs. There are no right or wrong answers. The definitions remain visible at the top of every screen.

Format for Each Item

Box A7. Item-screen template: primary classification, correspondence rating, and overlap ratings.

Statement [X] of 22
[Item text displayed here]
Q1. Primary Classification
Which construct does this statement best represent? [Response options randomised per participant]
○ Perceived Cognitive Assistance (PCA) – Helps my thinking
○ Perceived Usefulness (PU) – Improves results
○ Perceived Ease of Use (PEOU) – Easy to operate
○ Trust in the LLM – Reliable/safe
○ Trading Self-Efficacy (TSE) – My own skill
○ Other / None of the above [Select if the statement is confusing, meaningless, or clearly does not fit ANY of the provided definitions]
Q2. Correspondence Rating (7-point scale)
How well does this statement match the definition you selected?
1 = Extremely poor match | 2 = Very poor | 3 = Somewhat poor | 4 = Adequate | 5 = Somewhat good | 6 = Very good | 7 = Extremely good match
Q3. Overlap Ratings
To what extent could this statement also fit each of the other constructs?
Format: For each construct NOT selected as primary, display 1–7 scale.
1 = No fit at all | 2 | 3 = Slight fit | 4 | 5 = Moderate fit | 6 | 7 = Strong fit
Overlap ratings were collected only for the non-selected constructs; the overlap row for the construct chosen in Q1 was not shown, so no overlap rating exists for the selected construct by design.
Optional qualitative feedback about ambiguity was collected after the sort-and-rate task (see Section 7).
[Open text]

Attention Checks

Researcher-only (not shown to participants): Implementation note: Two attention checks are presented as separate screens inserted at randomised positions within the item sequence. These are not included in the 20 content items (16 PCA candidates + 4 fillers) and do not enter content-validation metrics. Participants thus see 22 total classification screens: 20 content items + 2 attention checks. Additionally, the four filler items function as implicit attention checks—systematic misclassification of obvious comparator items signals inattentive responding.

Attention Check 1: 'Data quality check: please select Perceived Usefulness (PU) for the Primary Classification for this item.'

Attention Check 2: 'To confirm attention, please select Trust in the LLM for this item and rate correspondence as 6 or 7.'

Section 6. Item Pool (20 Items Total: 16 PCA Candidates + 4 Fillers)

Researcher-only (not shown to participants): These 20 items are the content stimulus set analysed (16 PCA candidates + 4 fillers); two additional attention-check screens are presented separately and are excluded from content-validity metrics. All items are presented in randomised order without IDs. Judges see only the statement text. Revisions to PCA items address PEOU contamination (removing 'easy/clear' language), Trust contamination (reframing error-detection as user agency), and PU contamination (removing outcome/benefit framing). Participant-facing content (shown to participants): Judges saw only the statement texts below, presented one per screen in randomised order. Item IDs are included here for reporting and analysis only and were not shown in the survey.

PCA Candidate Items (16 Items)

Table S1. PCA candidate items presented to judges (16 items).

ID	Item Text (Final Revised)
PCA1	Using an LLM provides a structured path from my initial trading idea to a concrete order.
PCA2	Using an LLM helps me break complex trades into manageable steps.
PCA3	LLM support enables me to translate my market view into a precise, executable trade plan.
PCA4	LLM support helps me structure multi-leg or conditional orders in a more organised way.
PCA5	With LLM help, I can process more information at once without feeling overloaded.
PCA6	LLM support helps me filter and isolate the specific information required for a trading decision.
PCA7	LLM support helps me connect information from different sources into a coherent picture.
PCA8	Using an LLM enhances my ability to spot inconsistencies or gaps in my trading plans.
PCA9	LLM support helps me notice when numbers, dates, or assumptions in my plan do not line up.
PCA10	LLM support prompts me to scrutinize my trade for potential failure points before I commit capital.
PCA11	LLM support helps me compare alternative tactics for the same view side by side.
PCA12	LLM support helps me think through what-if scenarios and plan for different market paths.
PCA13	LLM support helps me keep my bearings and stay organised when markets move quickly.
PCA14	LLM support helps me understand the rationale behind a strategy's suitability for my market view.
PCA15	LLM support expands the range of trading scenarios I am able to mentally simulate.
PCA16	LLM support facilitates reflection on my decision-making process during trade review.

Comparator Construct Filler Items (4 Items)

Researcher-only (not shown to participants): Methodological justification: Following Anderson and Gerbing (1991), content validation tasks should include items from multiple constructs rather than presenting only focal-construct items. This reduces demand characteristics (the implicit expectation that 'all items should be classified as PCA'), increases task realism, and provides internal validation that judges can correctly classify unambiguous comparator items. One filler item per comparator construct (PU, PEOU, Trust, TSE) is included, deliberately written to be clear exemplars of their respective constructs. These fillers also function as implicit attention checks—systematic misclassification of obvious comparator items signals inattentive responding.

Researcher-only (not shown to participants): Presentation: Filler items are intermixed with PCA items in the randomised sequence. Judges see 20 total items (16 PCA candidates + 4 fillers). Filler item classifications are reported separately and do not affect PCA item retention decisions, but serve as validation that judges understood and applied the definitions correctly.

Table S2. Comparator construct filler items presented to judges (4 items).

Target	Filler Item Text
PU	Using an LLM for trading saves me time and improves my overall results.
PEOU	The LLM interface is straightforward and I find it easy to type my prompts and read the responses.
Trust	I believe the LLM provides accurate and reliable information that I can depend on.
TSE	I am confident in my own ability to analyse market conditions and execute trades successfully.

Researcher-only (not shown to participants): Expected classification rates: If definitions are clear and judges are attentive, filler items should achieve psa ≥ 0.85 for their target construct. Classification rates below 0.70 for any filler item suggest either (a) definitional ambiguity requiring refinement, or

(b) systematic inattention in the judge sample. Filler classification rates are reported as a quality indicator alongside Fleiss' kappa.

Researcher-only (not shown to participants): Revision summary: PCA1 changed from 'feels clearer' to 'provides a structured path' (removes PEOU trigger). PCA3 changed to 'enables me' (removes 'easier'). PCA6 changed to 'filter and isolate' (removes value language). PCA8 changed to 'enhances my ability' (removes 'effective'). PCA10 changed to 'prompts me to scrutinize' (reframes from passive 'identify' to active user agency—addresses Trust contamination). PCA14 simplified syntax. PCA15 changed to 'expands the range of scenarios I can mentally simulate' (removes novelty/benefit framing that bleeds into PU). PCA16 changed to 'facilitates reflection... during trade review' (anchors to tool use, reduces TSE bleed).

Section 7. Debrief and Completion

[FINAL SCREEN]

Box A8. Final screen: debrief and completion items.

Thank you for taking part.

This study examines how different survey statements about AI tools are interpreted by retail-type traders. Your classifications and fit ratings will help refine the PCA item pool before it is used in a larger trader survey.

Difficulty ratings (mandatory)

D1. How difficult was it to distinguish between Perceived Cognitive Assistance and Perceived Usefulness?

1 = Very easy | 2 | 3 = Moderate | 4 | 5 = Very difficult

D2. How difficult was it to distinguish between Perceived Cognitive Assistance and Trust?

1 = Very easy | 2 | 3 = Moderate | 4 | 5 = Very difficult

D3. How difficult was it to distinguish between Perceived Cognitive Assistance and Perceived Ease of Use?

1 = Very easy | 2 | 3 = Moderate | 4 | 5 = Very difficult

D4. Which categories felt most similar, and why? [Open text; optional]

[Open text box]

D5. Were any specific statements confusing or ambiguous? If so, please describe which ones and why. [Open text; optional]

Completion

Questions: Contact [PI email]

Completion Code: [Prolific completion code]

Section 8. Analysis Framework and Decision Rules (Researcher Reference Only)

Primary indices are computed at the item level using N eligible judges. This framework aligns with Colquitt et al. (2019) guidelines for definitional correspondence and definitional distinctiveness, with thresholds specified a priori.

8.1 Primary Indices

Proportion of substantive agreement (psa):

$psa = n_c / N$

where n_c is the number of judges who selected PCA as the primary classification, and N is the total number of eligible judges.

Substantive validity coefficient (csv):

$csv = (n_c - n_o) / N$

where n_o is the largest number of classifications of the same item into any single non-PCA construct (PU, PEOU, Trust, or TSE). Classifications into 'Other/None' are excluded from n_o but reduce n_c.

Mean correspondence rating (MCR):

MCR = Mean Q2 rating (1–7 scale) among judges who selected PCA as the primary classification.

Note: To compute Colquitt et al.'s normalized htc for benchmarking: htc = MCR / 7. This yields a 0–1 scale directly comparable to Colquitt's Table 5 norms (median htc = .88 across 112 published scales).

For decision rules below, we use raw MCR (1–7) for interpretability.

Definitional distinctiveness indices (htd_proximal and htd_plus_TSE):

Following Colquitt et al. (2019), distinctiveness is assessed via signed differences between intended and orbiting correspondence ratings. Because Trading Self-Efficacy (TSE) uses a different referent (baseline capability independent of tools) than the other constructs (all situated in LLM-assisted trading), two indices are computed:

Here, a is the number of response categories for correspondence/overlap ratings (a = 7), so (a − 1) = 6.

htd_proximal (primary index for retention decisions):

htd_proximal = Mean[(Intended correspondence − Orbiting correspondence)] / (a − 1)

computed using only the three *referent-matched, high-risk comparators*: PU, PEOU, and Trust. These constructs share the LLM-assisted trading context and represent the primary contamination risks identified in the manuscript.

htd_plus_TSE (supplementary index, reported but not decisive):

htd_plus_TSE = Mean[(Intended correspondence − Orbiting correspondence)] / (a − 1)

computed using all four comparators (PU, PEOU, Trust, TSE). This supplementary index confirms that PCA items are not collapsing into generic 'I am a skilled trader' statements, but is not used for retention thresholds because TSE's different referent makes distinctiveness comparatively easier to achieve.

Calculation steps (same procedure for both indices):

1. Restrict to the subset of judges who classified the item as PCA (i.e., n_c judges).
2. For these judges only: compute MCR as the mean of their Q2 (correspondence to PCA) ratings.
3. For these same judges only: compute mean Q3 ratings for each relevant orbiting construct.
4. For htd_proximal: compute three signed differences (MCR − Mean_PU), (MCR − Mean_PEOU), (MCR − Mean_Trust), then average and divide by 6.
5. For htd_plus_TSE: compute four signed differences (adding MCR − Mean_TSE), then average and divide by 6.

Rationale: This approach keeps intended and orbiting ratings on the same respondent base. Using htd_proximal for retention decisions ensures distinctiveness evidence is anchored to constructs where leakage is genuinely plausible, while htd_plus_TSE provides supplementary confirmation that items are not generic self-confidence statements.

Interpretation: Both indices range from −1 to +1. Positive values indicate the item corresponds more strongly to PCA than to orbiting constructs among judges who classified it correctly. For htd_proximal, values near zero suggest problematic overlap with PU, PEOU, or Trust specifically.

Methodological note on TSE referent:

Trading Self-Efficacy is intentionally operationalised as baseline capability independent of tools, rather than sharing the 'while using an LLM' referent of other constructs. This design choice tests

whether PCA items are non-redundant with general trading competence beliefs. Following Colquitt et al.'s (2019) guidance that different referents can inflate apparent distinctiveness, TSE is treated as a secondary comparator: included in overlap ratings and htd_plus_TSE reporting, but excluded from the primary htd_proximal index used for retention decisions. This approach preserves analytical rigour while avoiding the practically unjudgeable boundary that would result from forcing respondents to separate 'my ability when using tools' from 'assistance provided by tools.'

Construct-specific overlap indices:

For each comparator construct, compute the mean Q3 overlap rating (1–7) across judges who classified the item as PCA. Report separately:

- **Proximal overlaps (retention-decisive):** Mean_PU, Mean_PEOU, Mean_Trust
- **TSE overlap (supplementary):** Mean_TSE

This separation ensures that distinctiveness evidence used for retention decisions is anchored to constructs where leakage is genuinely plausible given shared referent context.

8.2 Decision Rules (A Priori – Calibrated per Psychometric Audit)

CORE item:

- psa ≥ 0.65 *(raised from 0.50 to require near two-thirds agreement)*
- MCR ≥ 5.25 (i.e., ≥ 75% of maximum; equivalent to htc ≥ .75)
- csv > 0.20 *(raised from > 0 to require meaningful margin over strongest competitor)*
- htd_proximal > 0.10 *(uses only referent-matched comparators: PU, PEOU, Trust)*
- No proximal-construct overlap > 4.5 *(i.e., Mean_PU, Mean_PEOU, and Mean_Trust each ≤ 4.5)*

Action: Retain.

BORDERLINE item:

- psa in [0.50, 0.64] OR MCR in [4.5, 5.24]
- csv > 0
- htd_proximal > 0 (positive but below .10)
- No proximal-construct overlap > 5.0

Action: Review wording; candidate for revision. Review post-task open-ended feedback (D4–D5) for clarity problems and perceived overlap explanations.

PROBLEMATIC item:

- psa < 0.50 OR MCR < 4.5 OR csv ≤ 0 OR htd_proximal ≤ 0
- OR any proximal-construct overlap > 5.0 *(overlap exclusion trigger for PU, PEOU, or Trust)*

Action: Rewrite substantially or drop.

HIGH-CONFIDENCE CORE (secondary label):

- psa ≥ 0.75 AND htd_proximal ≥ 0.20 AND all proximal overlap ratings ≤ 3.5

Action: Candidates for marker item designation in Wave-1 EFA.

8.3 Overlap Exclusion Rule

Even if an item achieves high psa and MCR, it may still be theoretically contaminated if judges perceive strong overlap with a competing construct. This rule operationalises the principle that high agreement can coincide with high perceived overlap, which would undermine the simple-structure assumption required for subsequent CFA.

Proximal constructs (retention-decisive): PU, PEOU, and Trust share the LLM-assisted trading referent with PCA and represent genuine contamination risks. Overlap thresholds for these constructs directly affect retention decisions.

- **Hard exclusion:** Mean Overlap > 5.0 on any proximal construct triggers PROBLEMATIC classification regardless of other indices.

- **Soft flag:** Mean Overlap > 4.0 on any proximal construct triggers close inspection of post-task open-ended feedback (D4–D5) and consideration for targeted rewording, even if other indices meet CORE thresholds.

TSE (supplementary, reported but not decisive): Because TSE uses a different referent (baseline capability independent of tools), high TSE overlap does not trigger retention decisions. Mean_TSE is reported for transparency, and unusually high TSE overlap (> 4.0) is noted as a potential signal that an item reads as generic self-confidence rather than tool-assisted cognition. However, this does not automatically downgrade item status.

8.4 Coverage Safeguard

After applying decision rules, retained items are checked to ensure that the item set preserves conceptual coverage across the PCA subdimensions (C1–C5). If a subdimension has no retained CORE item, the BORDERLINE item with the strongest MCR and highest htd is retained provisionally to preserve content coverage.

8.5 A Priori Contamination Risk Items

Items PCA8–PCA10 address error/gap detection and may show elevated Trust overlap despite revisions. Items PCA6 and PCA15 were revised to remove outcome/benefit framing but should be monitored for PU overlap. These items are retained intentionally at this stage to allow the naive-judge procedure to quantify definitional distinctiveness. This approach aligns with Colquitt et al. (2019) guidance on detecting construct contamination through content validation.

8.6 Inter-Rater Reliability

Report Fleiss' kappa for the categorical classification (Q1) across all 16 items as a global measure of inter-rater agreement. Kappa > .40 indicates moderate agreement; > .60 indicates substantial agreement.

8.7 Robustness Checks

Re-run analyses excluding participants who failed any attention check (AC1 or AC2).
Stratify results by the D1–D3 difficulty ratings to assess whether perceived construct similarity predicts misclassification patterns.
No comprehension-based robustness checks are conducted because first-attempt comprehension responses are not stored in the analysis export.