Supplementary Material D. Naïve-judge content-validation results.

Full protocol analysis with duplicate exclusion (first-submission rule).

Analysis Date: January 5, 2026 | Valid Judges: N = 48

D.1 Summary

Abbreviations: AC1 and AC2 = attention checks; IQR = interquartile range; SD = standard deviation; psa = proportion of substantive agreement; csv = construct substantive validity; MCR = mean correspondence rating; κ = Fleiss' kappa.

- Valid judges: N = 48 (after excluding one duplicate submission under the first-submission rule).
- Attention check pass rates: AC1 = 96.1%, AC2 = 98.0%, Joint = 96.1%
- Filler accuracy: PU = 81.2% (below 85% threshold), PEOU = 91.7%, Trust = 89.6%, TSE = 95.8%
- Overall PCA detection: Mean psa = 0.674, Median psa = 0.667
- Item retention: 7 CORE, 9 BORDERLINE, 0 PROBLEMATIC
- Change from N = 49: PCA5 reclassified from CORE to BORDERLINE (psa 0.653 to 0.646).
- Subdimension coverage: C1, C4, C5 covered; C2 and C3 require safeguard retention
- Fleiss' kappa (κ) = 0.011 for PCA items (Fleiss, 1971; interpretation: slight agreement; Landis & Koch, 1977).

Data quality note: duplicate submission and screening inconsistency

One duplicate submission was identified. Following a pre-specified first-submission rule, only the first submission was retained. The first submission indicated ineligibility on the screening item about AI chatbot use for trading, so the duplicate was excluded from analysis. Excluding this case reduced the analysis sample from N = 49 to N = 48 and moved item PCA5 from CORE to BORDERLINE (psa from 0.653 to 0.646).

D.2 Methodological Overview

This study follows the Colquitt et al. (2019) content validation protocol, which assesses whether survey items are correctly interpreted as representing their intended construct.

Constructs Under Evaluation

- PCA (Perceived Cognitive Assistance): The focal construct. Perception that using a large language model (LLM) scaffolds the trader's reasoning process.
- PU (Perceived Usefulness): Belief that using the LLM leads to better outcomes. The primary contamination risk.
- PEOU (Perceived Ease of Use): Belief that the LLM interface is easy to operate.
- Trust: Confidence that the LLM's outputs are accurate and reliable.
- TSE (Trading Self-Efficacy): Confidence in one's own baseline trading ability, independent of any tool.

Table S8 shows the number of participants retained at each stage of data cleaning. The final step excludes one duplicate submission following a pre-specified first-submission rule.

Table S8. Sample Flow

Step	N
Initial exported records	166
Remove preview/test records	154
Completed survey	154

Reached item-classification block	51
Pass attention check 1 (AC1)	49
Pass attention check 2 (AC2)	50
Pass both attention checks (pre-deduplication)	49
Exclude duplicate submission (first-submission rule)	48
Valid judges (final analysis sample)	48

The excluded duplicate involved inconsistent responses to the eligibility screen; exclusion follows the rule described in D.1.

Timing and demographic characteristics of the valid judge sample (N = 48). IQR = interquartile range; SD = standard deviation. Table S9 reports timing and demographic characteristics of the valid judge sample (N = 48). IQR = interquartile range; SD = standard deviation.

Table S9. Participant Descriptives

Variable	Summary
Duration: Median	1330 seconds (22.2 minutes)
Duration: IQR	1032 - 1876 seconds
Duration: Range	691 - 3081 seconds
Speeders (<5 minutes)	0 (none identified)
LLM users for trading	48/48 (100%)

Prolific Platform Validation

Demographic characteristics were obtained from Prolific where available. Among retained judges with available profile data (46/48), all reported investment experience. The mean age was 36.3 years (standard deviation (SD) = 11.2; N = 46) and 82.6% were male (N = 46). Two retained participants' profile demographics were unavailable due to consent revocation.

Panel A: Attention Checks

Attention checks verify that participants read carefully (Table S10.A) . Rates calculated on base sample (N=51) before AC filtering.

Table S10.A. Attention Checks

Attention Check	Pass N / Total	Pass Rate
AC1: 'Please select Perceived Usefulness (PU)'	49 / 51	96.1%
AC2: 'Please select Trust AND rate 6 or 7'	50 / 51	98.0%
Joint: Passed BOTH attention checks	49 / 51	96.1%

Panel B: Filler Item Accuracy

Filler items are 'obvious' exemplars of each comparator construct. Target accuracy: ≥85%. Calculated on valid judges (N=48). Table S10.B summarises filler-item accuracy for each comparator construct, calculated on valid judges (N = 48).

Table S10.B Filler Item Accuracy

Filler Item	Accuracy	Correct/N	Interpretation
PU: 'saves time, improves results'	81.2%	39/48	Below threshold - PU boundary difficult
PEOU: 'interface is straightforward'	91.7%	44/48	Above threshold
Trust: 'accurate and reliable'	89.6%	43/48	Above threshold

TSE: 'confident in my own ability'	95.8%	46/48	Above threshold

The PU filler item performed below the 85% threshold, which suggests that distinguishing PCA from perceived usefulness (PU) is difficult, consistent with the conceptual proximity of the two constructs.

D.3 Key Metrics: Definitions and Interpretation

This section defines all metrics used in Table S11. Understanding these is essential for interpreting retention decisions.

Primary Classification Metrics

psa (Proportion of Substantive Agreement): Proportion of judges who classified the item as PCA. Formula: psa = n_PCA / N. Range: 0–1. Higher is better. Threshold: ≥0.65 for CORE (Anderson & Gerbing, 1991).

csv (Construct Substantive Validity): Margin between PCA classifications and the strongest competing construct. Formula: csv = (n_PCA − n_max_competitor) / N. Range: −1 to 1. Threshold: >0.20 for CORE (Anderson & Gerbing, 1991).

MCR (Mean Correspondence Rating): Among PCA-classifiers, average rating of fit (1-7 scale). Reflects confidence. Threshold: ≥5.25 for CORE.

Top Competitor: Non-PCA construct receiving most classifications. Identifies contamination source.

Distinctiveness Metrics

Overlap Ratings (Ovlp_PU, Ovlp_PEOU, Ovlp_Trust): Among PCA-classifiers, mean rating of fit to other constructs (1-7). High overlap (>4.5) indicates dual-loading. Computed only among judges who selected PCA.

htd_proximal (Heterotrait Distinctiveness): Composite index: htd = mean(MCR − Overlap_k) / 6, where k ∈ {PU, PEOU, Trust}. TSE excluded (different referent). Range: -1 to 1. Threshold: >0.10 for CORE.

Decision rules for item status

CORE — Retain unchanged. ALL criteria must be met:

- psa ≥ 0.65
- MCR ≥ 5.25
- csv > 0.20
- htd_proximal > 0.10
- All proximal overlaps ≤ 4.5

BORDERLINE — Revise wording:

- Fails one or more CORE criteria but not PROBLEMATIC

PROBLEMATIC — Drop or rewrite:

- psa < 0.50 OR MCR < 4.5 OR csv ≤ 0 OR htd ≤ 0 OR any overlap > 5.0

Items sorted by status (CORE first), then by psa descending. Threshold violations are summarised in the 'BORDERLINE triggers' line below Table S11 and can be verified by comparison with the thresholds listed above. Table S11 reports the item-level content-validation results for the 16 PCA candidate items, including definitional correspondence and distinctiveness indices.

Table S11. Item-level content-validation results (N = 48; psa, csv, MCR, overlap ratings, and htd_proximal are defined in D.3).

Item	Sub	psa	csv	MCR	htd_proximal	Ovlp_PU	Ovlp_PEOU	Ovlp_Trust	Status
PCA12	C4	0.854	0.792	6.22	0.446	4.24	2.76	3.63	CORE
PCA15	C5	0.833	0.750	5.95	0.425	4.28	2.62	3.30	CORE
PCA3	C1	0.792	0.708	5.87	0.379	4.32	2.76	3.71	CORE
PCA14	C5	0.708	0.583	6.00	0.438	4.29	2.18	3.65	CORE
PCA1	C1	0.688	0.583	6.03	0.416	4.33	2.88	3.39	CORE
PCA11	C4	0.688	0.583	6.45	0.493	4.24	2.85	3.39	CORE
PCA16	C5	0.688	0.542	6.15	0.451	4.15	2.73	3.45	CORE
PCA7	C2	0.750	0.667	5.97	0.360	4.56	3.17	3.72	BORDERLINE
PCA5	C2	0.646	0.500	6.45	0.470	4.26	3.39	3.26	BORDERLINE
PCA8	C3	0.646	0.438	6.13	0.452	4.16	2.32	3.77	BORDERLINE
PCA4	C1	0.604	0.417	5.86	0.364	4.72	3.07	3.24	BORDERLINE
PCA6	C2	0.604	0.396	5.97	0.412	4.31	2.79	3.38	BORDERLINE
PCA9	C3	0.604	0.438	6.14	0.418	4.59	2.59	3.72	BORDERLINE
PCA2	C1	0.583	0.396	6.46	0.506	4.25	3.07	2.96	BORDERLINE
PCA13	C4	0.583	0.417	6.11	0.446	4.29	2.54	3.46	BORDERLINE
PCA10	C3	0.521	0.292	6.08	0.411	4.56	2.44	3.84	BORDERLINE

BORDERLINE triggers: PCA7 (Ovlp_PU>4.5); PCA5,8,4,6,9,2,13,10 (psa<0.65); PCA4,9,10 also have Ovlp_PU>4.5.

If a subdimension has no CORE items, retain the best BORDERLINE item to preserve content breadth. Table S12 summarises subdimension coverage and shows where the coverage safeguard was applied.

Table S12. Subdimension coverage (CORE = meets all criteria; BORDERLINE = needs targeted revision).

Code	Domain	CORE	BORDERLINE	Status	Action
C1	Structural scaffolding	2	2	Covered (≥1 CORE item)	Retain CORE
C2	Cognitive load management	0	3	No CORE items (safeguard retained)	Retain PCA5 (safeguard)
C3	Error/bias mitigation	0	3	No CORE items (safeguard retained)	Retain PCA8 (safeguard)
C4	Scenario navigation	2	1	Covered (≥1 CORE item)	Retain CORE
C5	Metacognitive learning	3	0	Covered (≥1 CORE item)	Retain CORE

Coverage safeguard applied: C2 (cognitive load) and C3 (error and bias) have no CORE items. To preserve content coverage, retain PCA5 (highest MCR in C2 = 6.45) and PCA8 (strong MCR with lower PU overlap in C3).

Inter-Rater Reliability

Fleiss' κ (inter-rater reliability)

Fleiss' kappa (κ) measures inter-rater agreement for categorical data with multiple raters (Fleiss, 1971). Range: −1 to +1, with 0 indicating chance agreement.

Interpretation benchmarks (Landis & Koch, 1977):

- κ < 0.20: Slight agreement
- κ 0.21-0.40: Fair agreement
- κ 0.41-0.60: Moderate agreement
- κ 0.61-0.80: Substantial agreement

D.4 Results

Table S13 reports Fleiss' κ for the PCA-only item set and for the full item set including fillers.

Table S13. Fleiss' κ for item sets

Item Set	Fleiss' κ
16 PCA candidate items only	0.011 (Slight)
20 items including fillers	0.316 (Fair)

The low κ = 0.011 for PCA items likely reflects (1) skewed marginal distributions (most items classified as PCA) and (2) deliberately proximal comparator constructs. When filler items are included, κ increases to 0.316, which suggests that judges distinguish constructs when items are unambiguous.

To support interpretation of the retained-item set, we report (i) item-level correspondence and boundary diagnostics for the nine retained PCA items (Table S14) and (ii) a concise annotation of each retained item's targeted cognitive mechanism and process-only wording features (Table S15).

Table S14. Retained-item correspondence and boundary diagnostics (valid judges N = 48; retained PCA-9)

Item	Sub	Status	psa	nPCA	n_PU_misclass	nmax_other	Ovlp_PU	Ovlp_PEOU	Ovlp_Trust
PCA1	C1	CORE	0.688	33	5	5	4.33	2.88	3.39
PCA11	C4	CORE	0.688	33	5	5	4.24	2.85	3.39
PCA12	C4	CORE	0.854	41	3	3	4.24	2.76	3.63
PCA14	C5	CORE	0.708	34	6	6	4.29	2.18	3.65
PCA15	C5	CORE	0.833	40	4	4	4.28	2.62	3.30
PCA16	C5	CORE	0.688	33	4	7	4.15	2.73	3.45
PCA3	C1	CORE	0.792	38	4	4	4.32	2.76	3.71
PCA5	C2	BORDERLINE	0.646	31	7	7	4.26	3.39	3.26
PCA8	C3	BORDERLINE	0.646	31	10	10	4.16	2.32	3.77

Table S15. Retained-item annotation table (process-only wording features)

Item	Macro-code	Targeted cognitive mechanism	Process-only wording features
PCA1	C1	Structural/path support	From idea to order sequencing; "structured path"; avoids outcome terms; decision-time framing

PCA3	C1	Translation and planning	Translate view into trade plan; "precise, executable plan"; no profitability/returns language; LLM as support
PCA5	C2 (safeguard)	Cognitive-load relief	More information without overload; affective load cue; no "better results" phrasing
PCA8	C3 (safeguard)	Error-checking and gap detection	Spot inconsistencies/gaps; verification support; no tool-reliability claims
PCA11	C4	Side-by-side comparison	Compare tactics; decision structuring; avoids evaluative outcome statements
PCA12	C4	What-if scenario planning	Scenario exploration; planning for different paths; process-level phrasing
PCA14	C5	Rationale and suitability understanding	Understand rationale/fit; learning support; avoids "improves outcomes" wording
PCA15	C5	Mental simulation breadth	Expands scenario simulation range; cognitive capability; no value appraisal terms
PCA16	C5	Reflection and review	Facilitates reflection during trade review; metacognitive support; not performance claim

To support interpretation of the low κ and to show where boundary difficulty concentrates, we report the full item-by-construct classification counts (including "Other/none") and the dominant competing classification for each retained PCA item (Table S16).

Table S16. Confusion summary: classification counts by construct (valid judges N = 48)

Item	PCA	PU	PEOU	Trust	TSE	Other_none
PCA1	33	5	3	4	2	1
PCA2	28	9	7	3	1	0
PCA3	38	4	3	1	2	0
PCA4	29	9	6	1	1	2
PCA5	31	7	4	3	2	1
PCA6	29	10	1	6	2	0
PCA7	36	4	4	4	0	0
PCA8	31	10	1	3	3	0
PCA9	29	7	1	8	2	1
PCA10	25	11	1	7	2	2

PCA11	33	5	5	1	3	1
PCA12	41	3	1	1	2	0
PCA13	28	8	5	4	2	1
PCA14	34	6	1	4	3	0
PCA15	40	4	1	0	3	0
PCA16	33	4	1	3	7	0
Filler_PU	1	39	4	1	3	0
Filler_PEOU	0	1	44	2	1	0
Filler_Trust	0	4	1	43	0	0
Filler_TSE	2	0	0	0	46	0

Table S17 summarises the dominant misclassification pattern for each retained PCA item by identifying the primary competing construct.

Table S17. Item-level misclassification profiles (valid judges N = 48)

Item	Primary_class	Primary_count	Top_competing_class	Top_competing_count	Other_none_count
PCA1	PCA	33	PU	5	1
PCA3	PCA	38	PU	4	0
PCA5	PCA	31	PU	7	1
PCA8	PCA	31	PU	10	0
PCA11	PCA	33	Tie: PU / PEOU	5	1
PCA12	PCA	41	PU	3	0
PCA14	PCA	34	PU	6	0
PCA15	PCA	40	PU	4	0
PCA16	PCA	33	TSE	7	0

D.5 Recommendations

CORE Items — Retain Unchanged (7 items)

PCA1, PCA3, PCA11, PCA12, PCA14, PCA15, PCA16

BORDERLINE Items — Targeted Revision (9 items)

High PU overlap (>4.5):

- PCA7 (Ovlp_PU=4.56): Replace 'coherent picture' with 'structured outline'
- PCA4 (Ovlp_PU=4.72): Reframe from 'organised way' to 'specify legs in a structured plan'
- PCA9 (Ovlp_PU=4.59): Strengthen agency: 'prompts me to check internal consistency'
- PCA10 (Ovlp_PU=4.56): Replace 'failure points' with 'audit internal consistency'

Low psa (<0.65):

- PCA5 (psa=0.646, C2 safeguard): Minor revision—emphasize cognitive process over capacity
- PCA8 (psa=0.646, C3 safeguard): Add 'active scrutiny' language for user agency
- PCA6 (psa=0.604): Emphasize cognitive-load mechanism over efficiency
- PCA2 (psa=0.583): Replace 'manageable steps' with 'explicit reasoning steps'
- PCA13 (psa=0.583): Anchor to 'orderly decision process' not market outcomes

Recommended PCA-9 item set for Wave 1 exploratory factor analysis (EFA)

Proceeding to Wave 1 pilot exploratory factor analysis (EFA) with: 7 CORE items plus two safeguard items (PCA5 for C2 coverage and PCA8 for C3 coverage) for a total of 9 items.

- Wave 1 target sample size: n approximately 60 (minimum 5:1 participant-to-item ratio).
- Wave 2 target sample size: n = 200–300, stratified by experience and LLM usage.
- Include perceived usefulness (PU) marker items for discriminant validity assessment.