

Developing an AI-Generated Language-Specific Epistemic Justification Inventory: A Preliminary Validation Using a Machine-Translation Error-Detection Task

Jinru Luo, Xinting Zhang

Southern University of Science and Technology, Shenzhen 518055, Guangdong, China

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: Foreign-language learners now treat generative AI as a routine source of information about how a language works, even though what these tools produce is fluent yet not always correct. How learners decide whether to trust such output (in epistemic terms) has therefore become a practical concern for language education. Adapting Bråten and colleagues' Internet-Specific Epistemic Justification Inventory (ISEJI), we wrote a 12-item inventory covering three theorized ways of justifying AI-generated language information: against one's own knowledge, by cross-checking several sources, and by appeal to authority. Eighty-four non-language-major undergraduates completed it alongside a machine-translation (MT) error-detection task. Internal consistency was reasonable for a first draft (overall $\alpha = .79$; subscales .68–.75). An exploratory factor analysis recovered the personal-justification items cleanly but folded the multiple-sources and authority items together, hinting that in the AI setting learners may distinguish only between internal and external grounds for belief. More usefully, total scores told apart students who caught the translation error from those who missed it (4.98 vs. 4.65; $t = 2.08$, $p = .041$; $d = 0.47$), and a one-SD rise in justification raised the odds of a successful catch by roughly 58% (OR = 1.58). We read these results as encouraging but preliminary: the sample is small, the structure is unsettled, and the criterion is thin. The inventory is offered as a starting point, not a finished scale.

Keywords: Epistemic justification; AI-assisted language learning; Machine translation; Questionnaire development; Criterion validity

Online publication: June 26, 2026

1. Introduction

When a student needs to phrase something in English today, the answer increasingly comes from ChatGPT rather than from a teacher, a dictionary, or a search engine^[1]. What returns is different in kind from a web-page answer. It arrives as confident, well-formed prose with no visible author and no link to follow, and because a language model is optimized to sound right, it is becoming harder to detect the problems. A learner

who accepts such output uncritically is exposed precisely where the tool is weakest. The question this study takes up is how learners judge whether AI-generated language information deserves their trust.

That judgment is what the literature on epistemic cognition calls *epistemic justification*: the grounds a person invokes when deciding that a claim is believable. Bråten and colleagues moved the idea online, arguing that people justify knowledge on the Internet in partly distinct ways, and built the ISEJI to capture three of them: checking a claim against one's own knowledge and reasoning (personal justification), corroborating it across several sources (justification by multiple sources), and leaning on an authoritative or expert source (justification by authority) ^[2]. The inventory held together well with Norwegian preservice teachers ^[2], and Internet-specific justification has been linked to deeper online learning ^[3].

What the field lacks is a comparable instrument for AI-generated content, let alone one anchored in a subject such as language. The gap is not merely administrative. AI output is not epistemically interchangeable with a web page: the web has publishers and traceable links, whereas a model collapses author, source, and authority into a single fluent voice. Strategies that make sense against a website, such as “see whether other sites agree” and “check for an expert source”, may not carry over, and may not even remain separable, once the source is a chatbot. Interest in how students trust tools like ChatGPT is growing ^[4], but measurement has not kept pace.

Language learning is a good place to test the idea. Learners depend on AI for word choice, grammar, and translation, yet these are exactly the areas where machine translation and language models remain imperfect and can produce subtle errors ^[5], for instance, with culture-bound terms, technical vocabulary, or shifts of register. If epistemic justification is a genuine disposition rather than a questionnaire artifact, learners who report justifying AI information should also be the ones who notice when a translation goes wrong. A demonstrated link of that sort would do double duty: it would speak to the instrument's validity and to the broader claim that epistemic justification shows up in what people actually do.

We therefore adapted the ISEJI into an inventory for AI-generated language information and conducted a first empirical test with Chinese foreign-language learners. The framing throughout is deliberately modest. With 84 respondents, this is a pilot; the structural findings are exploratory, and the confirmatory work (cross-validation and invariance testing) is left for a larger study.

2. Background and research questions

2.1. Epistemic justification, from print to the web

Epistemic justification names the standards a reader brings to bear on whether a claim can be believed. Early studies located it in reading across single and multiple texts and treated sophisticated justification as a matter of not trusting one source, weighing evidence across texts, and sizing up where a claim comes from ^[6]. Bråten and colleagues then made the case that justification is to some degree tied to setting: how one vets a knowledge claim online need not match how one vets a printed page. Their Internet-specific version split the disposition into personal, multiple-sources, and authority justification, a three-part structure that has since been reproduced and tied to information vetting ^[7].

2.2. Why AI-generated content is not just more of the web

Carrying the construct from the web to AI output involves as much rupture as continuity. The continuity is easy to state: evaluating a claim is still evaluating a claim, and all three strategies remain available in

principle. One can weigh an answer against what one already knows, compare answers across tools, or ask whether the model is any good at this sort of question. The rupture is in what “source” and “authority” now mean. A web page borrows authority from an institution or a named expert; a model’s “authority” is something closer to its own competence and reliability, as in judging whether an answer came from a stronger, steadier system. “Multiple sources,” likewise, can mean several websites or several chatbots checked against one another. Because these meanings drift, whether the ISEJI’s tidy three-way split survives the move to AI is an empirical matter, not a foregone conclusion and, as it turns out, worth doubting.

2.3. Justification as behavior, not just attitude

A recurring claim in this literature is that epistemic beliefs are supposed to do something, to surface in how people learn and reason, not merely in how they answer a questionnaire ^[8-9]. In the case of AI-generated language information, a matching behavioral measure is readily available: an MT error-detection task asks the reader to justify a translation against the source, against their own sense of the language, against an alternative rendering. That is justification enacted rather than reported. If inventory scores predict who catches the error, the instrument earns a piece of criterion validity that self-report measures rarely secure.

2.4. Research questions

Three questions guide this study:

- (1) Items and reliability. Do the items discriminate, and are the overall and subscale reliabilities acceptable?
- (2) Structure. Does the inventory recover the theorized three dimensions?
- (3) Criterion validity. Do scores separate students who detect an MT error from those who do not?

3. Method

3.1. Participants

The respondents were non-language-major undergraduates in the 2025 cohort at a STEM university in China, giving 84 usable cases (ages 19–20). All had about one year’s experience using AI tools (ChatGPT, DeepSeek, and the like) for language study and had the proficiency assumed by the translation task (CET-4). Participation was voluntary and with consent.

3.2. Instruments

- (1) The inventory. We adapted the ISEJI ^[2] by swapping its “web information” frame for “information about language usage produced by AI tools” and rewording items for language learning. Twelve items are rated 1 (strongly disagree) to 7 (strongly agree), four to each theorized dimension:
- (2) Personal justification (Q1, Q4, Q7, Q10): testing AI information against one’s own knowledge and logic; e.g., “I evaluate whether this information is consistent with my own understanding or with relevant knowledge.”
- (3) Multiple-sources justification (Q2, Q5, Q8, Q11): corroborating across platforms, tools, or sites; e.g., “I compare several answers from different AI tools.” and “I compare AI’s answer with other external sources.”
- (4) Authority justification (Q3, Q6, Q9, Q12): weighing the source’s competence or standing; e.g., “I try to determine whether it comes from a more dependable and robust large-language model.”

- (5) The error-detection task (criterion). To get a behavioral anchor, we used a Chinese-to-English error-detection task. Several culture-heavy Chinese passages (on the Tao Te Ching, shadow puppetry, traditional Chinese medicine, bridge building, and the guqin) were each rendered into English by a language model (ChatGPT-4, accessed December 2, 2024) and by a neural MT system (Google Translate, accessed December 2, 2024). Respondents judged whether the translation contained errors (compulsory task) and labeled the error type (optional task). Scored against a key, each student's result was coded 1 (caught an error) or 0 (did not). The criterion used here is whether the student caught the error at all.

3.3. Procedure

Both parts were done on a paper questionnaire in one 30-minute sitting, inventory first on the front side of the paper, then the task on the back side of the paper. The questionnaire was administered anonymously; responses carried only a sequential code, linked to the task by that code. The first author ran the session. Respondents were allowed to use tools to check words or expressions only when they did the error detection task.

3.4. Analysis

Analyses ran in Python (pandas, numpy, scipy, scikit-learn): item statistics (means, SDs, skew, kurtosis, and corrected item-total correlations(CITCs)); Cronbach's α overall and by subscale; factorability by KMO and Bartlett's test; factor retention by the eigenvalue-over-one rule, parallel analysis, and theory, with principal-axis factoring and varimax rotation; inter-dimension correlations; and, for criterion validity, point-biserial correlations, Welch t tests, Cohen's d, and a binary logistic regression of detection on justification.

4. Results

4.1. Items and descriptives

No values were missing, and all stayed within the 1–7 range. **Table 1** gives the item statistics. The pattern of means is itself informative: personal-justification items sat highest (Q1, Q7, Q10 between 5.50 and 5.58), authority items lowest (Q6 and Q9 both 4.19). Read plainly, these students lean on their own knowledge when vetting AI and are far more reluctant to defer to a tool's supposed expertise, a sensible instinct, arguably, given how unreliable such deference can be. CITCs mostly ran .34–.59, but Q4 and Q7 came in at .24, and both, along with Q1 and Q10, were sharply left-skewed and peaked (Q7: skew -1.76 , kurtosis 3.60). The personal items, in other words, are bumping against the ceiling.

Table 1. Item statistics and corrected item-total correlations (N = 84)

Item	Dimension	M	SD	Skew	Kurtosis	CITC
Q1	Personal	5.50	1.01	-1.71	2.78	.34
Q4	Personal	5.30	1.23	-0.99	0.44	.24
Q7	Personal	5.52	1.02	-1.76	3.60	.24
Q10	Personal	5.58	1.00	-1.32	2.25	.35
Q2	Multiple sources	4.45	1.46	-0.17	-1.12	.41
Q5	Multiple sources	4.90	1.50	-0.77	-0.33	.46

Item	Dimension	M	SD	Skew	Kurtosis	CITC
Q8	Multiple sources	4.56	1.50	−0.61	−0.90	.59
Q11	Multiple sources	4.75	1.34	−0.75	−0.44	.58
Q3	Authority	4.25	1.54	−0.13	−1.19	.55
Q6	Authority	4.19	1.52	0.00	−1.13	.42
Q9	Authority	4.19	1.44	0.05	−1.01	.47
Q12	Authority	4.71	1.39	−0.24	−0.78	.51

4.2. Reliability

The full inventory returned $\alpha = .79$, which is respectable for a first version. Subscale values were .68 (personal), .75 (multiple sources), and .71 (authority). The personal subscale slipping under .70 is no surprise once one notices it carries the two weak, ceiling-bound items (Q4, Q7). With only four items per subscale, none of this is alarming (Table 2).

Table 2. Overall and subscale reliability (N = 84)

Scale	Items	Cronbach's α
Full inventory	12	.79
Personal justification	4	.68
Multiple-sources justification	4	.75
Authority justification	4	.71

4.3. Exploratory factor analysis

Factorability was adequate if unspectacular: KMO = .66, Bartlett's $\chi^2(66) = 291.7$, $p < .001$. The leading eigenvalues were 3.73, 1.80, 1.15, 1.02, 0.85, and so on. The eigenvalue-over-one rule would keep four factors, but with the third and fourth barely clearing 1.0, that rule is doing little more than counting noise. Parallel analysis was stricter: only the first two observed eigenvalues beat the 95th percentile of random data ($3.73 > 1.85$; $1.80 > 1.61$; the third, 1.15, fell short of 1.43). The data support two factors.

For comparability with the ISEJI, we still report the theory-guided three-factor solution (varimax; 44.3% of variance, itself on the low side). Table 3 tells a clear story. The four personal items load together and nowhere else (.42–.83). The multiple-sources and authority items, by contrast, refuse to separate: Q2, Q5, Q8, Q11, Q3, Q9, and Q12 pile onto one factor, and only Q6 breaks away to define a lonely third factor about whether the AI is “capable.” Whatever the three-item theory says, these respondents seem to draw one line: between checking against themselves and checking against something outside the model, not two.

Table 3. Rotated loadings, principal-axis factoring (N = 84; varimax; loadings $\geq .40$ in bold)

Item	Theorized	Factor 1 (external)	Factor 2 (personal)	Factor 3 (AI capability)	h^2
Q1	Personal	.25	.62	−.10	.46
Q4	Personal	−.03	.50	.15	.28
Q7	Personal	.04	.42	.11	.19
Q10	Personal	.09	.83	.03	.70
Q2	Multiple sources	.45	.08	.19	.25
Q5	Multiple sources	.59	.03	.15	.37

Item	Theorized	Factor 1 (external)	Factor 2 (personal)	Factor 3 (AI capability)	h ²
Q8	Multiple sources	.70	.14	.19	.54
Q11	Multiple sources	.77	.10	.12	.62
Q3	Authority	.44	.15	.42	.39
Q6	Authority	.13	−.01	.94	.90
Q9	Authority	.40	.12	.35	.30
Q12	Authority	.37	.21	.37	.32
% variance		18.4	13.5	12.3	Cum. 44.3

4.4. Correlations among dimensions

The subscale correlations say the same thing more compactly (**Table 4**). Personal justification barely moves with the other two ($r = .21, .23$), while multiple-sources and authority track each other at .54. That is the two-factor picture again, arriving by a second route.

Table 4. Dimension and total-score correlations ($N = 84$)

	Personal	Multiple	Authority	Total
Personal	1.00			
Multiple sources	.21	1.00		
Authority	.23	.54	1.00	
Total	.56	.83	.83	1.00

4.5. Criterion validity

Forty-five students (53.6%) caught the translation error, and 39 (46.4%) did not, close to a coin flip, a sobering result in its own right. Split by outcome (**Table 5**), the detectors scored higher on the full inventory (4.98 vs. 4.65): $r_{pb} = .23$; Welch's $t = 2.08$, $p = .041$; $d = 0.47$. Every subscale leaned the same way ($d = 0.32$ – 0.37), but none reached significance on its own, which is about what one should expect from $n = 84$ split in two.

A logistic regression of detection on the three standardized subscales gave positive coefficients throughout (personal OR = 1.34, authority OR = 1.25, multiple sources OR = 1.16) and an in-sample AUC of 0.62 (modest discrimination). With the standardized total as the sole predictor, OR = 1.58: a one-SD gain in reported justification bought roughly 58% better odds of catching the error. Reported justification, then, does appear to be reflected in behavior; but the predictive signal comes from students' overall level of justification rather than from any one of the three strategies.

Table 5. Inventory scores by detection outcome ($N = 84$)

Dimension	Detected ($n = 45$) $M \pm SD$	Missed ($n = 39$) $M \pm SD$	r_{pb}	t	p	d
Personal	5.61 $\pm$ 0.59	5.33 $\pm$ 0.91	.18	1.64	.106	0.37
Multiple sources	4.83 $\pm$ 0.91	4.48 $\pm$ 1.26	.16	1.43	.158	0.32
Authority	4.52 $\pm$ 1.10	4.13 $\pm$ 1.03	.18	1.67	.099	0.36
Total	4.98 $\pm$ 0.62	4.65 $\pm$ 0.83	.23	2.08	.041	0.47

5. Discussion

Three findings deserve attention, though each carries a limitation.

The items work well enough, with one soft spot. An α of .79 and subscales in the .68–.75 band show that measuring epistemic justification for AI content is feasible, and the construct did not dissolve on contact with the new setting. The soft spot is the personal subscale, dragged down by Q4 and Q7 and by a ceiling that leaves little variance to work with. Two readings seem plausible. One is substantive: “I check it against what I know” is such a flattering, low-cost thing to endorse that almost everyone does, which is a warning about social desirability as much as about item quality. The other is technical: some items are quietly double-barreled (Q1 asks about “my own understanding *or* relevant knowledge”), and a respondent who agrees with one half but not the other has nowhere to go. Either way, the personal items are the first thing a revision should fix.

The structure is not the one we assumed, and that is the interesting part. Personal justification held up as a factor; multiple sources and authority did not stay apart. Both the factor analysis and the plain correlations point instead to a single external dimension. This is easy to rationalize after the fact, and the rationalization is offered here for what it is worth: online, “check other sites” and “find an expert” are different acts, but with AI they blur, since comparing chatbots and asking which model is stronger both amount to looking outside a single AI answer for confirmation. If that reading is right, justification is not a fixed three-part trait but something that regroups with the information source, consistent with earlier arguments that the construct is context-sensitive^[6]. With 84 respondents and parallel analysis calling for two factors, the honest summary is narrower: personal justification is distinct, while the multiple-sources/authority distinction did not hold up. Only a larger confirmatory study can tell whether the two-factor structure is real or just an artifact of this small sample.

The link to behavior is the paper’s best evidence and still slim. That total scores separate the students who caught the error from those who missed it ($d = 0.47$; $OR = 1.58$) is a trustworthy result, because it ties a self-report to something the respondent did rather than said^[8]. It is also the result most easily oversold. The subscale differences, though all in the right direction, were not significant, so there is no warrant for claiming that one strategy detects errors better than another.

6. Limitations and next steps

The limits are real and they shape what can be claimed.

The sample is the first. Eighty-four respondents support exploration and nothing sturdier; a KMO of .66 and 44% of variance explained are passing marks, not good ones. The obvious next move is a larger, independent sample ($N \geq 300$) with a confirmatory factor analysis pitting the two-factor form against the three, followed by invariance tests across sex and major.

The criterion is the second. One binary outcome wastes information that was there for the taking: how precisely an error was located, whether its type was named correctly, and it leaves the instrument’s wider validity untouched. Placing the inventory in a network of related measures would show what it converges with and what it does not^[1,10,11].

The third limitation is self-report. The personal items show a ceiling effect. Almost everyone endorses them, which may reflect social desirability, or simply that “I check AI against what I already know” is an easy, flattering thing to affirm. Anonymous administration should have blunted such pressure, but self-report

alone cannot rule it out. Future work should add a social-desirability check and, where possible, corroborate self-reports with behavioral measures.

Wording is the last. A few items discriminate poorly, and one (Q8) still speaks of “other websites,” a holdover from the web-era instrument that fits the AI frame awkwardly.

None of this is fatal; it is the ordinary distance between a pilot inventory and a scale. What the study delivers is a usable draft with early evidence that it measures something real, and a fairly specific list of what to repair before the confirmatory round.

7. Conclusion

We adapted the ISEJI into a short inventory for AI-generated language information and conducted a first test with foreign-language learners. Reliability was acceptable; personal justification emerged as its own dimension while multiple-sources and authority merged into a broader external one; and total scores predicted who caught a real translation error. Overall, the findings recommend the inventory as a starting instrument for studying and eventually teaching how learners judge what AI tells them about language, but only if the next study is larger and includes better validity anchors.

Funding

The Southern University of Science and Technology (Project No.: XJZLGC202429)

Disclosure statement

The author declares no conflict of interest.

References

- [1] Xu J, 2025, Harnessing AI in Academic Writing: The Complex Interplay of AI Literacy and Self-Directed Learning Among University L2 Learners. *System*, 135: 103855.
- [2] Bråten I, Brandmo C, Kammerer Y, 2019, A Validation Study of the Internet-Specific Epistemic Justification Inventory with Norwegian Preservice Teachers. *Journal of Educational Computing Research*, 57(4): 877–900.
- [3] Bråten I, Strømsø HI, Samuelstuen MS, 2005, The Relationship Between Internet-Specific Epistemological Beliefs and Learning Within Internet Technologies. *Journal of Educational Computing Research*, 33(2): 141–171.
- [4] Rind IA, Bhatti J, Chellappan K, 2026, From Blind Trust to Critical Inquiry: Epistemic Beliefs and Student Engagement with ChatGPT in Higher Education. *Jurnal Kejuruteraan*.
- [5] Horbach A, Pehlke J, Laarmann-Quante R, Ding Y, 2023, Crosslingual Content Scoring in Five Languages Using Machine-Translation and Multilingual Transformer Models. *International Journal of Artificial Intelligence in Education*.
- [6] Bråten I, Britt MA, Strømsø HI, Rouet JF, 2011, The Role of Epistemic Beliefs in the Comprehension of Multiple Expository Texts: Toward an Integrated Model. *Educational Psychologist*, 46(1): 48–70.
- [7] Ståhl T, 2024, Justification for Knowing in a Digitalised Landscape: An Empirical Exploration. *Journal of Digital Social Research*, 6(3): 59–77.

- [8] Maggioni L, Parkinson MM, 2008, The Role of Teacher Epistemic Cognition, Epistemic Beliefs, and Calibration in Instruction. *Educational Psychology Review*, 20(4): 445–461.
- [9] Guilfoyle L, McCormack O, Erduran S, 2024, The Role of Research Experiences in Developing Pre-Service Teachers' Epistemic Beliefs. *Teaching and Teacher Education*, 144: 104599.
- [10] Ng DTK, Wu W, Leung JKL, Chiu TKF, Chu SKW, 2023, Design and Validation of the AI Literacy Questionnaire. *British Journal of Educational Technology*, 55(3): 1082–1104.
- [11] Carolus A, Koch MJ, Straka S, et al., 2023, MAILS – Meta AI Literacy Scale. *Computers in Human Behavior: Artificial Humans*, 1(2): 100014.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.