Reviewer's Addendum (Post-Revision)

Overview

This document synthesizes feedback from three automated reviews of the revised paper "Piensa Again: Testing the Foreign Language Effect in LLMs via LoRA Adapters." The reviews examined the paper from cognitive science, methods/reproducibility, and epistemic hygiene perspectives.

The paper has addressed many issues from the previous review round. Remaining concerns focus on: (1) residual mechanistic language, (2) "replicating" framing in Abstract, and (3) task comprehension as a competing explanation.

Consensus Issues (Raised by Multiple Reviewers)

1. "Replicating" Language in Abstract

Source: CogSci Review, Epistemic Review

Current text: "replicating the directional pattern observed in Tversky & Kahneman (1981)"

Concern: "Replicating" implies reproducing the phenomenon. You observed the same directional asymmetry in a completely different system (LLMs vs humans).

Suggested revision: "showing the same directional asymmetry as reported in Tversky & Kahneman (1981)"

---

2. Task Comprehension as Primary Alternative Hypothesis

Source: CogSci Review (central concern), Methods Review (acknowledged)

Concern: The English adapter gradient (EN > ES > HE > ZH) might reflect task difficulty gradients rather than FLE-like processing: - Chinese dataset had 4,750 samples (vs 5,000 for others) - Hebrew and Chinese were GPT-4o translated, not human-verified - Mistral has English-dominant training data - The 15% unclear rate in ZH+HE suggests comprehension issues

Current treatment: Acknowledged in Limitations but should be foregrounded as a parallel interpretation.

Suggested action: Add to Analysis or Discussion a section treating task difficulty as an alternative explanation with equal weight to FLE interpretation.

---

3. Residual Mechanistic Language

Source: Epistemic Review (systematic issue)

Phrases implying internal processes not measured:

Current	Suggested
"decision heuristics"	"output patterns"
"comprehension failures"	"reduced response validity"
"confuse the model"	"reduce response consistency"
"interference"	"non-compositional effects"
"role-binding" (as mechanism)	"role-instruction prefix"

CogSci Review: Key Points

Strengths Noted

- Transparent about limitations
- Honest reporting of Spanish anomaly
- Methodological care with role-binding prefix
- Clear disclaimer about measuring choice frequencies, not cognitive biases

Remaining Concerns

1. **FLE Mischaracterization**: Costa et al. found *reduced* bias in L2, not just magnitude variation. All 16 conditions show positive framing effects; this differs from finding reduced bias.
2. **+62% Effect Size**: The HE+ES condition shows +62% framing effect, substantially larger than typical human effects (~20-30%). Without human baselines, magnitude interpretation is ungrounded.
3. **Spanish Anomaly as Disconfirming**: CogSci reviewer argues this should be treated as disconfirming evidence for the operationalization, not just a "challenge."
4. **L1/L2 Terminology**: Despite caveats, L1/L2 language throughout may mislead readers. Consider removing from title or reframing as "adapter-prompt alignment effects."

CogSci Verdict

> "This is an interesting computational experiment about adapter-prompt interactions, but the connection to human bilingual cognition is tenuous, and the results are more consistent with task difficulty confounds than with FLE-like processing."

Methods Review: Key Points

Assessment: Clear with Minor Clarifications

Strengths: - Comprehensive experimental design documentation - Transparent handling of evaluation artifacts - Honest reporting of anomalies - Low unclear rates enable confident interpretation

Remaining Clarifications Needed:

Item	Status
Training data quality control procedures	Needs clarification
Adapter convergence assessment methodology	Missing
Validation loss variation (0.93-1.17) interpretation	Needs clarification

Reproducibility Risks: 1. Training data quality variation (EN/ES human-verified vs HE/ZH LLM-translated) 2. Temperature=0.7 introduces sampling variance 3. Fixed 100 iterations without convergence criteria

Methods Verdict: > "A reader could re-run this experiment and understand which results are interpretable. The authors appropriately flag anomalous conditions."

Epistemic Review: Key Points

Assessment: Close to Excellent, Minor Revisions Needed

Strengths: - Abstract clearly scopes to "choice frequencies" not "cognitive biases" - Repeated acknowledgment of operationalization limitations - Discussion appropriately questions operationalization validity - Excellent limitation: "Without parallel human data…"

Required Revisions:

1. **Abstract**: Change "replicating" → "showing the same directional asymmetry as"
2. **Introduction**: Add hedging to Costa et al. citation:
    - Current: "bilinguals exhibited reduced cognitive biases"
    - Revised: "Costa et al. (2014) reported that, in their experiments, bilinguals exhibited…"

3. **Analysis**: Change mechanism claims to correlational:
    - Current: "if adapter training creates language-specific response tendencies"
    - Revised: "if adapter-prompt matching predicts larger effects"
4. **Discussion**: Replace mechanistic language:
    - "comprehension failures" → "reduced response validity"
    - "confuse the model" → "reduce response consistency"
    - "interference" → "non-compositional effects"
5. **Methods**: Consider "role-instruction prefix" instead of "role-binding" to describe instruction type rather than supposed mechanism.

Summary of Actionable Revisions

Critical (All Three Reviewers Agree)

☐ **Abstract**: "replicating" → "showing the same directional asymmetry as"

High Priority (Two Reviewers)

☐ **Foreground task comprehension** as alternative explanation with equal weight
☐ **Remove mechanistic language** throughout (see table above)

Medium Priority (Single Reviewer, Well-Argued)

☐ **Introduction**: Hedge Costa et al. citation ("in their experiments")
☐ **Analysis**: Change "creates" → "predicts" for adapter-effect correlations
☐ **Methods**: Consider "role-instruction prefix" terminology
☐ **Discussion**: Address +62% effect size interpretation (larger than human baselines)

Consider for Future Work

- Remove FLE/L1/L2 from title entirely
- Add explicit comprehension checks for probabilistic statements
- Include human baseline study on identical stimuli

Acknowledgments

This addendum was generated by synthesizing automated reviews from: - **CogSci Review**: anthropic/claude-sonnet-4.5 - **Methods Review**: anthropic/claude-sonnet-4 - **Epistemic Review**: anthropic/claude-sonnet-4.5

Review generation: 2026-01-01