

# JIAYI DENG, PH.D.

Applied Research Scientist | AI Fairness, Psychometrics & Human-AI Evaluation

New York Metro | etheldeng0726@gmail.com | [GitHub](#) | [Live Demo](#)

## PROFESSIONAL SUMMARY

---

Applied research scientist and psychometrician with 5+ years of experience evaluating high-stakes assessment and selection systems through validation, measurement invariance, reliability, equating, simulation, and behavioral research. Hands-on experience evaluating generative-AI-assisted expert-review workflows and building a deployed human-AI fairness audit with subgroup, intersectional, calibration, counterfactual, and human-override analyses. Advanced R expertise with applied Python and SQL; developer of reproducible research tools and co-author of eight peer-reviewed journal articles.

## CORE EXPERTISE

---

**AI fairness & human-AI evaluation:** subgroup and intersectional analysis, selection-rate/adverse-impact screening, equal opportunity/equalized odds, calibration, counterfactual testing, human overrides, behavioral evaluation, fairness scorecards

**Psychometrics & research:** measurement invariance, DIF/DDF, reliability, criterion-related validity, equating, IRT, signal detection theory, simulation, meta-analysis, cognitive labs, response-process evidence

**Technical:** R; Python (pandas, scikit-learn, Fairlearn, statsmodels, Streamlit); SQL/DuckDB; Plotly; pytest; GitHub Actions; reproducible pipelines; model and dataset documentation

## SELECTED AI FAIRNESS PROJECT

---

**Human-AI Fairness Audit Lab** | *Independent Applied Research Project*

[GitHub](#) | [Live Demo](#)

- Built and deployed a public, reproducible audit of a synthetic AI-assisted candidate-review workflow using Python, Fairlearn, scikit-learn, SQL/DuckDB, Plotly, and Streamlit; used separate synthetic training (6,000), validation (2,000), held-out audit (2,400), and monitoring (800) partitions for evaluation and demonstration, not real hiring decisions.
- Evaluated subgroup and intersectional selection outcomes, error rates, calibration, threshold sensitivity, counterfactual consistency, and bootstrap uncertainty; compared human-only, model-only, and AI-assisted review conditions.
- Analyzed human overrides, disagreement and escalation pathways, and automation-bias risk to assess when human review mitigates or amplifies model disparities across decision contexts.
- Created reusable synthetic datasets, 56 automated tests, artifact validation and a GitHub Actions CI workflow, plus a fairness scorecard, dataset and model cards, an audit report, a monitoring plan, and a public interactive dashboard.

## PROFESSIONAL EXPERIENCE

---

**Human Resources Research Organization (HumRRO)** | *Research Scientist*

2024–Present

- Conduct psychometric validation and independent quality reviews for statewide assessments, evaluating scaling and equating, score comparability, alignment, data quality, and documentation of limitations and residual risk.
- Contribute to the design and evaluation of generative-AI-assisted workflows for standard-to-standard, item-to-standard, and evidence-mapping tasks across assessment alignment studies.
- Evaluate chat-based and batch AI outputs within expert-review workflows, assessing output quality, consistency, scalability, documentation requirements, validation evidence, and implications for human oversight.
- Serve on the psychometric team for AAMC PREview, a high-stakes situational judgment assessment used in medical-school admissions; contribute weighted-score and subscore analyses, equating, psychometric evaluation, and hybrid signal-detection-theory scoring research.
- Conduct think-aloud cognitive labs and structured behavioral coding to evaluate response processes and identify potential sources of measurement error or unintended differences; translate findings into decision-ready technical reports for multidisciplinary stakeholders.

**University of Minnesota** | *Graduate Research Assistant*

2020–2024

- Designed empirical, simulation, and meta-analytic studies quantifying how differential rapid guessing affects reliability, score accuracy, equating, measurement invariance, and comparability across populations.
- Implemented IRT, DIF/DDF analysis, response-time thresholding, multigroup concurrent scaling, computerized adaptive testing, and reproducible parameter-estimation and diagnostic pipelines in R.
- Co-developed and maintained the CRAN R package *aberrance*, automating detection and filtering of anomalous test-taking behavior in large-scale assessment data; co-authored eight peer-reviewed journal articles.

## ADDITIONAL EXPERIENCE

---

**Pearson** | *Research Assistant*

2023–2024

- Performed psychometric analyses for aimswebPlus progress-monitoring assessments as part of an external evaluation by the National Center on Intensive Intervention.

**Pearson VUE** | *Psychometric Intern*

2023

- Designed and executed a simulation study evaluating test-wiseness in multiple-response items; generated and analyzed data and translated findings into an NCME conference proposal.

**Human Resources Research Organization (HumRRO)** | *Psychometric Intern*

2022

- Supported multi-state alignment studies through data management, R programming, test-map development, psychometric analysis, and quality control.

## SELECTED RESEARCH & PUBLICATIONS

---

- Can AI Motivate Students? A Systematic Review and Meta-analysis on AI's Role in Student Motivation and Engagement. Manuscript under review.
- Gorney, K., & Deng, J. (2024). aberrance: An R package for detecting aberrant behavior in test data. *Applied Psychological Measurement*.
- Rios, J. A., & Deng, J. (2024). Is effort-moderated scoring robust to multidimensional rapid guessing? *Educational and Psychological Measurement*.
- Rios, J. A., & Deng, J. (2024). A comparison of response-time threshold scoring procedures in mitigating bias from rapid guessing behavior. *Educational and Psychological Measurement*.
- Deng, J., & Rios, J. A. (2022). Investigating the effect of differential rapid guessing on population invariance in equating. *Applied Psychological Measurement*, 46(7), 589–604.
- Rios, J. A., & Deng, J. (2022). Quantifying the biasing effect of rapid guessing on reliability estimates: A simulation and meta-analytic investigation. *Applied Psychological Measurement*, 46(1), 40–52.

## SELECTED APPLIED RESEARCH CONTRIBUTIONS

---

- AI impact evidence synthesis: Co-authored a systematic review and meta-analysis examining how AI-enabled interventions affect student motivation and engagement across study designs, educational contexts, and outcome measures.
- High-impact selection and assessment systems: Contributed to AAMC PREview psychometric analyses, weighted-score and subscore equating, hybrid signal-detection-theory scoring, and independent reviews of statewide scaling and equating.
- Fairness and subgroup evaluation: Investigated differential rapid guessing and population invariance in equating; evaluated response-time procedures that mitigate bias from non-effortful responding; studied DIF/DDF and score comparability across groups.

## EDUCATION

---

<b>University of Minnesota</b>   <i>Ph.D., Quantitative Methods in Education; Minor in Statistics</i> Advisors: Mark Davison and Michael Rodriguez	<b>2024</b>
<b>University of Kansas</b>   <i>M.S., Research, Evaluation, Measurement and Statistics</i>	<b>2020</b>
<b>Beijing Normal University–Hong Kong Baptist University United International College</b>   <i>B.S., Applied Psychology</i>	<b>2018</b>

## SELECTED RECOGNITION

---

- Graduate Student Annual Teaching Award, University of Minnesota (2022–2023) • Hauge Graduate Fellowship (2022–2023) • Reviewer, *Educational and Psychological Measurement* and *Educational Measurement: Issues and Practice*