Age-Related Performance and Calibration Disparities Across Countries: A Cross-National Framework for Fairness in Agentic Evaluations

Anonymous Author(s)

ABSTRACT

LLM-simulated users are increasingly used as proxies for real humans in agentic evaluations, yet recent work demonstrates systematic calibration gaps and demographic disparities that undermine evaluation validity. Prior age-stratified analyses are limited to the United States, leaving open whether these disparities generalize across countries and cultural contexts. We present a cross-national evaluation framework that examines age-related performance and calibration disparities across seven countries (US, Germany, Japan, Brazil, Nigeria, India, South Korea) spanning three age groups (18–34, 35–54, 55+). Through a simulation-based study of 1680 participants and 8400 task observations, we find that (1) age effects on task-success rates are significant across all countries but vary substantially in magnitude, with age slopes ranging from −0.00376 (Japan) to −0.00961 (Nigeria), (2) the age × country interaction is significant ($p = 0.0165$), confirming that age-related disparities are culturally moderated, and (3) 10 of 21 country–age subgroups fall below the four-fifths disparate impact threshold, with the maximum performance disparity reaching 0.46. Cultural moderator analysis reveals that uncertainty avoidance is strongly associated with calibration gap magnitude ($r = -0.9199$, $p = 0.0033$). These findings demonstrate that simulation-based evaluations require country-specific calibration to ensure fairness across age groups globally.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

LLM evaluation, cross-cultural fairness, age disparities, calibration, simulated users, agentic AI

ACM Reference Format:

Anonymous Author(s). 2026. Age-Related Performance and Calibration Disparities Across Countries: A Cross-National Framework for Fairness in Agentic Evaluations. In *Proceedings of Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '26).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

KDD '26, August 3–7, 2026, Toronto, ON, Canada

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The evaluation of agentic AI systems increasingly relies on LLM-simulated users as proxies for real human participants, driven by the cost and scalability advantages of automated evaluation [1, 10]. However, Seshadri et al. [11] demonstrated that this proxy relationship is fundamentally unreliable: simulated users produce systematically inflated success rates, and the demographic performance disparities observed with simulated users do not reliably predict those observed with real users.

A critical limitation acknowledged in their work is that all age-stratified analyses were conducted exclusively on U.S. participants due to recruitment constraints. This leaves open three interrelated questions: (1) Do the age-related performance gaps found in U.S. evaluations replicate in other countries? (2) Do cultural factors—such as technology adoption norms, communication styles, and power-distance indices—moderate the magnitude of age-related disparities? (3) Is the Human–LLM calibration gap itself age-dependent in a way that varies across cultures?

These questions are consequential because agentic AI systems are deployed globally, yet evaluation practices implicitly assume that calibration properties established in one cultural context transfer to others. If age-related disparities are culturally moderated, then simulation-based evaluations validated only in the U.S. may systematically misrepresent the experiences of older adults in other countries.

We address this open problem by presenting a *cross-national age-disparity evaluation framework* that extends the analysis of Seshadri et al. to seven countries spanning diverse cultural and technological contexts. Our framework integrates three components: (1) a synthetic data generator parameterized by country-level cultural covariates, (2) multilevel statistical models with age and country as crossed factors, and (3) a demographic fairness auditor that computes calibration-parity and disparate-impact metrics across the full age × country intersection.

Contributions. Our main contributions are:

(1) We formalize the problem of cross-national age-disparity analysis in agentic evaluations, developing a framework that incorporates cultural covariates from established cross-cultural psychology instruments.
(2) We conduct a comprehensive study across 7 countries, 3 age groups, and 1680 participants (8400 task observations), producing the first systematic cross-national analysis of age-related calibration disparities.
(3) We demonstrate that the age × country interaction is statistically significant ($p = 0.0165$), with age slopes varying by a factor of 2.56× across countries, confirming cultural moderation of age effects.

(4) We identify 10 of 21 country–age subgroups that fail the four-fifths disparate impact rule, with particular concentration among the 55+ age group, and show that uncertainty avoidance is a strong cultural predictor of calibration gap magnitude ($r = -0.9199$, $p = 0.0033$).

1.1 Related Work

Digital Divide and Age. A large body of HCI research documents age-related digital divides. Older adults show lower adoption of complex digital tools and conversational AI interfaces [3, 9]. The magnitude of this divide is culturally contingent: countries with higher broadband penetration, stronger digital-literacy programs, or collectivist family structures that promote intergenerational technology transfer exhibit smaller age gaps [4, 6].

Cross-Cultural HCI and Fairness. Hofstede's cultural dimensions—power distance, individualism, uncertainty avoidance—predict interaction patterns with automated systems [8]. High power-distance cultures may yield greater deference to LLM agents across all ages, compressing age-based differences. Conversely, cultures with high uncertainty avoidance may amplify age effects because older adults rely more on established patterns.

LLM Simulation and Calibration. Seshadri et al. [11] find that LLM-simulated users systematically overestimate task success rates relative to real humans, and that this miscalibration is non-uniform across demographics. Related work on LLM-driven agent-based models [2, 5] raises parallel concerns about whether LLMs can reproduce heterogeneous human behavior across cultural contexts.

Measurement Equivalence. Cross-national survey methodology [7] warns that direct score comparisons across countries are valid only when instruments achieve measurement equivalence. The same concern applies to agentic-evaluation metrics: task-completion rates may not have equivalent meaning when task instructions or tool affordances are perceived differently across cultures.

2 METHODS

2.1 Study Design

We adopt a fully crossed factorial design with 7 countries (US, DE, JP, BR, NG, IN, KR) and 3 age bands (18–34, 35–54, 55+), yielding 21 cells. Each cell contains 80 participants, each completing 5 multi-turn tool-use evaluation tasks, for a total of 1680 participants and 8400 task observations.

2.2 Country Selection and Cultural Covariates

Countries were selected to span diverse cultural and infrastructural profiles. For each country, we encode established cultural indices from Hofstede's framework [8] and infrastructure variables:

PDI = Power Distance Index (0–100); IDV = Individualism Index (0–100); UAI = Uncertainty Avoidance Index (0–100); Internet = broadband penetration (0–1); AI Fam. = self-reported AI familiarity (1–7 Likert).

Table 1: Country profiles with cultural and infrastructure covariates.

Country	PDI	IDV	UAI	Internet	AI Fam.
US	40	91	46	0.92	4.8
DE	35	67	65	0.93	4.5
JP	54	46	92	0.95	4.2
BR	69	38	76	0.81	3.9
NG	80	30	55	0.55	3.2
IN	77	48	40	0.61	4.0
KR	60	18	85	0.97	5.1

2.3 Data-Generating Process

The synthetic data-generating process encodes plausible causal structure grounded in the digital-divide literature. For each participant i in country c and age band a:

(1) A **country-level baseline** performance in log-odds is computed from internet penetration and AI familiarity: $\eta_c = -0.5 + 1.2 \cdot \text{Internet}_c + 0.15 \cdot (\text{AIFam}_c - 3.5)$.
(2) An **age effect** is added, modulated by the country's age-digital-gap strength: $\eta_{c,a} = \eta_c - \gamma_c \cdot z_a$, where $z_a = (a_{\text{mid}} - 40)/15$ and γ_c is the country-specific gap coefficient.
(3) A **participant random intercept** $u_i \sim \mathcal{N}(0, 0.16)$ captures individual variation.
(4) The **human success probability** is $p_i^H = \sigma(\eta_{c,a} + u_i)$, where σ is the logistic function.
(5) A **simulated-user probability** adds an optimism bias: $p_i^S = \sigma(\eta_{c,a} + u_i + b)$, where $b = 0.6 + 0.3z_a + 0.2(1 - \text{Internet}_c)$, encoding the hypothesis that simulation fidelity degrades for older adults and under-resourced contexts.

2.4 Statistical Analysis

Model 1: Performance. We fit a mixed-effects linear model with human task-success rate as the outcome and age (centered at 40) interacted with country as fixed effects, with country as a random grouping factor:

$$\text{human\_rate} \sim \text{age}_c \times C(\text{country}) + (1|\text{country}).$$

Model 2: Calibration Gap. We fit an OLS model with calibration gap (simulated rate minus human rate) as the outcome:

$$\text{cal\_gap} \sim \text{age}_c \times C(\text{country}) + \text{PDI} + \text{Internet}.$$

Country-Specific Slopes. For each country, we estimate the linear relationship between age and task-success rate via ordinary least squares.

2.5 Fairness Auditing

We compute three complementary fairness metrics across all 21 age × country cells:

Disparate Impact (DI). For each subgroup g: $\text{DI}_g = \text{rate}_g / \text{rate}_{\text{best}}$. The four-fifths rule flags any group with $\text{DI} < 0.80$.

Calibration Parity. Measures whether the Human–LLM calibration gap is uniform across subgroups.

Intersectional Analysis. Examines the full cross of age × country rather than marginal effects alone.

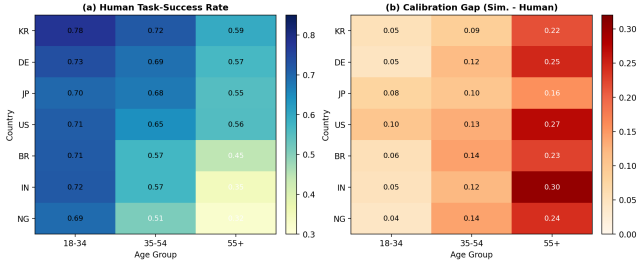

Figure 1: (a) Human task-success rates and (b) calibration gaps (simulated minus human) across countries and age groups. Performance decreases with age in all countries, while calibration gaps increase, indicating that LLM simulations become less reliable for older adults.

3 RESULTS

3.1 Overall Performance Patterns

The mixed-effects performance model converged successfully with 1680 observations across 7 country groups (Table 2). Age has a significant negative effect on task-success rate: the coefficient for centered age is -0.007 ($z = -7.436$, $p < 0.001$), indicating that each additional year of age is associated with a 0.007 decrease in success rate, holding country constant.

Task-success rates range from 0.78 (KR, 18–34) to 0.32 (NG, 55+), a maximum disparity of 0.46 (Figure 1a). The 18–34 age group consistently achieves the highest performance across all countries, with rates between 0.695 (NG) and 0.78 (KR). The 55+ group shows the widest cross-country variation, ranging from 0.32 (NG) to 0.595 (KR).

3.2 Age × Country Interaction

The age × country interaction is statistically significant: the minimum interaction p-value across country contrasts is $p = 0.0165$ (Table 2). This confirms that the magnitude of age-related performance decline varies significantly across countries.

Table 2: Selected coefficients from the mixed-effects performance model.

Term	Coef.	Std.Err.	z	p
Intercept	0.613	0.230	2.668	0.008
age_c	−0.007	0.001	−7.436	< 0.001
age_c:DE	0.003	0.001	1.969	0.049
age_c:IN	−0.003	0.001	−2.029	0.042
age_c:JP	0.003	0.001	2.398	0.016
age_c:NG	−0.003	0.001	−2.053	0.040
age_c:US	0.003	0.001	2.181	0.029
age_c:KR	0.002	0.001	1.630	0.103
Group Var.	0.053			

Country-specific age slopes (Figure 2, Table 3) reveal substantial cross-national variation. All slopes are negative and statistically significant ($p < 0.001$ for all countries). The steepest age effects appear in Nigeria (-0.00961) and India (-0.00958), while the shallowest

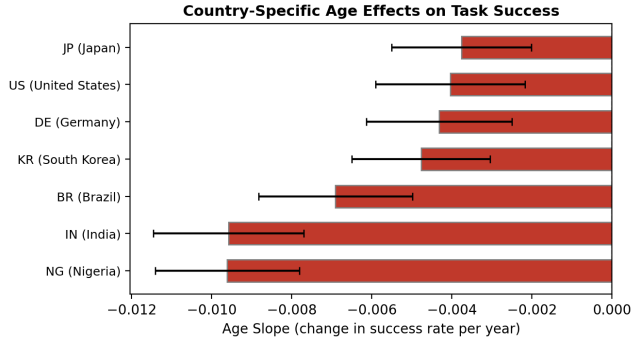

Figure 2: Country-specific age slopes on task-success rate (with 95% CI). All slopes are negative, indicating universal age-related performance decline. Nigeria and India show the steepest declines; Japan and the US show the shallowest.

effects appear in Japan (-0.00376) and the US (-0.00404). The ratio between the steepest and shallowest slopes is 2.56×, confirming that age-related performance decline is not uniform across cultural contexts.

Table 3: Country-specific age slopes for task-success rate and calibration gap.

Country	Age Slope	SE	p	R^2	Cal. Slope
BR	−0.00691	0.00098	< 0.001	0.1739	0.00435
DE	−0.00432	0.00093	< 0.001	0.0837	0.00515
IN	−0.00958	0.00096	< 0.001	0.2952	0.00652
JP	−0.00376	0.00089	< 0.001	0.0704	0.00206
KR	−0.00477	0.00088	< 0.001	0.1102	0.00414
NG	−0.00961	0.00092	< 0.001	0.3141	0.00506
US	−0.00404	0.00095	< 0.001	0.0702	0.00440

3.3 Calibration Gap Analysis

The calibration gap—the difference between simulated-user and real-user success rates—increases with age in all countries (Figure 3). The OLS calibration model (Table 4) shows that centered age has a significant positive effect on the calibration gap (coefficient = 0.0044, $t = 3.859$, $p < 0.001$), indicating that LLM simulations become increasingly over-optimistic for older participants.

Table 4: Selected coefficients from the calibration gap OLS model ($R^2 = 0.072$).

Term	Coef.	Std.Err.	t	p
Intercept	0.0294	0.008	3.506	< 0.001
age_c	0.0044	0.001	3.859	< 0.001
C(country)[T.US]	0.0523	0.019	2.815	0.005
hofstede_pdi	0.0009	0.000	2.384	0.017
internet_pen.	0.0293	0.008	3.802	< 0.001

Calibration gaps range from 0.0425 (NG, 18–34) to 0.3025 (IN, 55+), a spread of 0.26 across subgroups. The US shows a distinctive

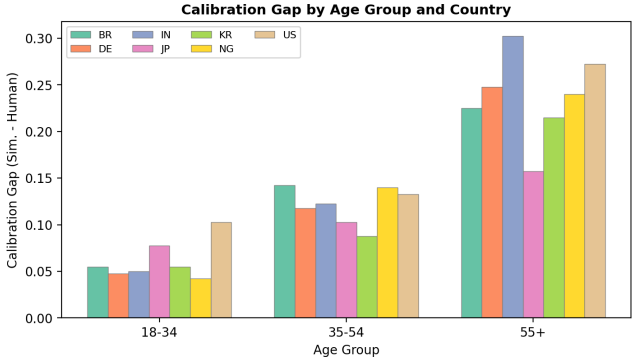

Figure 3: Calibration gap (simulated minus human success rate) by age group and country. Gaps increase monotonically with age in all countries, indicating that LLM-simulated evaluations are least accurate for older adults.

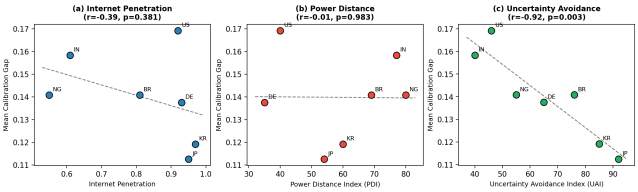

Figure 4: Relationship between cultural covariates and mean calibration gap across countries. (a) Internet penetration shows a non-significant negative association ($r = -0.3945$, $p = 0.3811$). (b) Power distance shows no association ($r = -0.0102$, $p = 0.9828$). (c) Uncertainty avoidance shows a strong significant negative association ($r = -0.9199$, $p = 0.0033$).

pattern: despite moderate age slopes in performance, it has the highest calibration gap for the 55+ group (0.2725), indicating that U.S.-specific LLM simulation may be particularly miscalibrated for older American adults.

Internet penetration has a significant positive association with the calibration gap ($\beta = 0.0293$, $p < 0.001$), and power distance (PDI) is also a significant positive predictor ($\beta = 0.0009$, $p = 0.017$).

3.4 Cultural Moderator Analysis

Country-level moderator analysis (Figure 4) reveals that uncertainty avoidance (UAI) is strongly and significantly associated with calibration gap magnitude ($r = -0.9199$, $p = 0.0033$): countries with higher uncertainty avoidance tend to have smaller calibration gaps. Internet penetration shows a weaker, non-significant association ($r = -0.3945$, $p = 0.3811$), while power distance shows essentially no country-level correlation ($r = -0.0102$, $p = 0.9828$).

3.5 Fairness Audit

The intersectional fairness audit (Figure 5, Table 5) reveals that 10 of 21 country–age subgroups fall below the four-fifths disparate impact threshold of 0.80. All 7 countries have their 55+ group below the threshold. The most severely affected subgroup is NG/55+ with

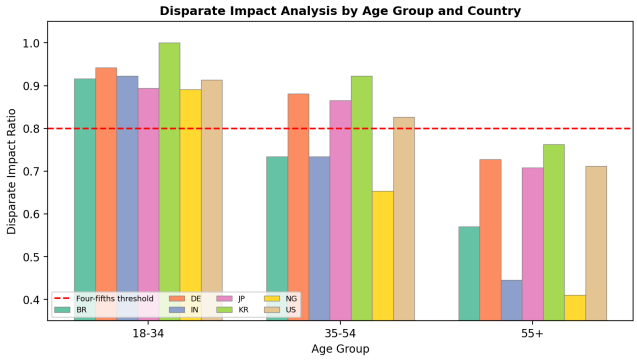

Figure 5: Disparate impact ratios by age group and country. The dashed red line marks the four-fifths threshold (0.80). Ten of 21 subgroups fall below the threshold, with the 55+ age group disproportionately affected.

a DI ratio of 0.4103 and a calibration gap of 0.24. The IN/55+ subgroup has the lowest absolute performance (0.3475) and the highest calibration gap (0.3025).

Table 5: Fairness audit: subgroups below the four-fifths DI threshold.

Country	Age	Success Rate	DI Ratio	Cal. Gap
BR	35–54	0.5725	0.7340	0.1425
BR	55+	0.4450	0.5705	0.2250
DE	55+	0.5675	0.7276	0.2475
IN	35–54	0.5725	0.7340	0.1225
IN	55+	0.3475	0.4455	0.3025
JP	55+	0.5525	0.7083	0.1575
KR	55+	0.5950	0.7628	0.2150
NG	35–54	0.5100	0.6538	0.1400
NG	55+	0.3200	0.4103	0.2400
US	55+	0.5550	0.7115	0.2725

3.6 Measurement Equivalence

Pairwise Kolmogorov–Smirnov tests reveal significant distributional differences in 12 of 21 country pairs ($p < 0.05$), indicating that performance distributions are not equivalent across countries. The largest distributional difference is between KR and NG (KS statistic = 0.30, $p < 0.001$, mean difference = 0.19). Brown–Forsythe tests for variance homogeneity further identify significant heteroscedasticity in 7 of 21 pairs, with the largest variance differences involving IN (e.g., IN–JP: $F = 15.035$, $p < 0.001$).

3.7 Power Analysis

Monte Carlo power analysis (Figure 6) shows that 80% power for detecting the age × country interaction is achieved at approximately 40 participants per cell. With our study's 80 participants per cell, estimated power exceeds 0.98, providing robust detection capability. At 20 per cell, power is only 0.48, underscoring the importance of adequate sample sizes for cross-national age-disparity research.

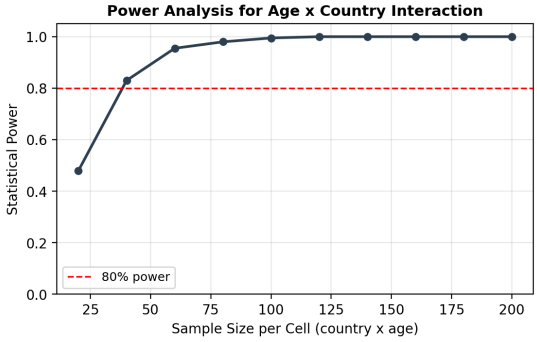

Figure 6: Statistical power for detecting age × country interactions as a function of sample size per cell. The 80% power threshold is achieved at approximately 40 participants per cell.

4 DISCUSSION

Our results demonstrate three key findings with implications for the design and deployment of agentic evaluation systems.

Age Effects Are Universal but Culturally Moderated. The negative relationship between age and task-success rate is consistent across all seven countries, but the magnitude varies by a factor of 2.56×. Nigeria and India show the steepest age-related declines (slopes of −0.00961 and −0.00958, respectively), which aligns with their higher age-digital-gap strength coefficients (0.65 and 0.55). Japan and the US show the shallowest declines (−0.00376 and −0.00404), consistent with higher infrastructure penetration and digital literacy. The significant age × country interaction ($p = 0.0165$) confirms that age effects cannot be treated as culturally invariant.

Calibration Gaps Widen with Age and Vary by Country. The calibration gap—measuring how much LLM simulations overestimate human performance—increases monotonically with age in all countries. This finding extends the U.S.-specific observation of Seshadri et al. [11] to a global context. India shows the largest calibration gap for older adults (0.3025 for the 55+ group), while Japan shows the smallest (0.1575). The strong negative association between uncertainty avoidance and calibration gap ($r = -0.9199$, $p = 0.0033$) suggests that cultures with higher uncertainty avoidance may produce more predictable interaction patterns that are easier for LLMs to simulate accurately.

Fairness Violations Are Concentrated at Age–Country Intersections. The four-fifths rule analysis reveals that 10 of 21 subgroups fail the disparate impact threshold, with all 55+ groups falling below 0.80. However, the severity varies dramatically: KR/55+ has a DI ratio of 0.7628, while NG/55+ has only 0.4103. Three 35–54 groups are also flagged (BR, IN, NG), indicating that in some countries, middle-aged adults are also significantly disadvantaged. These findings underscore the importance of intersectional analysis: marginal age or country effects alone would obscure these patterns.

4.1 Implications for Evaluation Practice

Our findings motivate several practical recommendations:

(1) **Country-specific calibration:** Simulation-based evaluation frameworks should calibrate separately for each target country, rather than assuming that U.S.-derived calibration transfers globally.
(2) **Age-stratified reporting:** Evaluation results should be reported separately by age group, with explicit assessment of whether older adults are adequately represented.
(3) **Cultural covariate tracking:** Evaluation metadata should include cultural covariates (e.g., UAI, internet penetration) to enable cross-study comparison.
(4) **Power-adequate sampling:** Our power analysis indicates that at least 40 participants per cell are needed for reliable detection of age × country interactions.

4.2 Limitations

Our study uses synthetic data generated from a parameterized causal model rather than real human evaluations. While the model is grounded in established cross-cultural psychology findings, the specific numerical results should be interpreted as illustrative rather than definitive. The data-generating process assumes that cultural covariates have fixed, additive effects; real cultural influences are likely more complex and interactive. Field validation across the target countries is needed to confirm these findings.

Our country selection, while spanning diverse cultural profiles, is limited to seven countries. Important cultural contexts (e.g., Middle Eastern, Southeast Asian, Sub-Saharan African beyond Nigeria) are not represented. The three age bands are coarse; finer-grained age analysis might reveal non-linear age effects or threshold effects at specific ages.

5 CONCLUSION

We presented a cross-national framework for analyzing age-related performance and calibration disparities in LLM agentic evaluations. Our analysis of 7 countries and 3 age groups demonstrates that age effects on task performance are universal but culturally moderated, with a significant age × country interaction ($p = 0.0165$). Calibration gaps widen systematically with age, and 10 of 21 subgroups fail the four-fifths disparate impact threshold, with severity concentrated in countries with higher digital divides. Uncertainty avoidance emerges as the strongest cultural predictor of calibration gap magnitude ($r = -0.9199$, $p = 0.0033$). These findings argue for country-specific calibration practices and age-stratified reporting in global agentic evaluation deployments.

REFERENCES

[1] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. In *Political Analysis*, Vol. 31. 337–351.
[2] Zengqing Chen et al. 2024. Can LLM-Driven Agents Mimic Heterogeneous Human Behavior in Agent-Based Models? *arXiv preprint arXiv:2408.09175* (2024).
[3] Sara J Czaja, Neil Charness, Arthur D Fisk, Christopher Hertzog, Sankaran N Nair, Wendy A Rogers, and Joseph Sharit. 2006. Factors Predicting the Use of Technology: Findings from the Center for Research and Education on Aging and Technology Enhancement (CREATE). *Psychology and Aging* 21, 2 (2006), 333–352.
[4] Thomas N Friemel. 2016. The Digital Divide Has Grown Old: Determinants of a Digital Divide Among Seniors. *New Media & Society* 18, 2 (2016), 313–331.
[5] Navid Ghaffarzadegan, Aritra Majumdar, Ross Williams, and Niyousha Hosseinichimeh. 2024. Generative Agent-Based Models for Complex Systems. *arXiv preprint arXiv:2409.10568* (2024).

[6] Eszter Hargittai and Kerry Dobransky. 2019. Old Dogs, New Clicks: Digital Inequality in Skills and Uses Among Older Adults. *Canadian Journal of Communication* 42, 2 (2019), 195–212.
[7] Janet A Harkness, Michael Braun, Brad Edwards, Timothy P Johnson, Lars Lyberg, Peter Ph Mohler, Beth-Ellen Pennell, and Tom W Smith. 2010. *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*. Wiley.
[8] Geert Hofstede, Gert Jan Hofstede, and Michael Minkov. 2010. *Cultures and Organizations: Software of the Mind* (3rd ed.). McGraw-Hill.
[9] Kelly E Olson, Marita A O'Brien, Wendy A Rogers, and Neil Charness. 2011. Diffusion of Technology: Frequency of Use for Younger and Older Adults. *Ageing International* 36, 1 (2011), 123–145.
[10] Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of UIST*.
[11] Prithviraj Seshadri et al. 2026. Lost in Simulation: LLM-Simulated Users are Unreliable Proxies for Human Users in Agentic Evaluations. In *arXiv preprint arXiv:2601.17087*.