Beyond Accuracy: A Replication Fidelity Framework for Trustworthy LLM Evaluation in Social Science Applications

Anonymous authors
Paper under double-blind review

Abstract

Current LLM evaluation approaches do not always detect systematic biases that undermine trustworthy deployment in social science applications. Using family ideal vignettes from established surveys across China (n=5,186) and the United States (n=5,906), we systematically compared five state-of-the-art LLMs against human responses using a novel Replication Fidelity Index (RFI). Our counter-intuitive finding reveals that successful replication requires capturing human disagreement and variability rather than just central tendencies—a fundamental limitation missed by standard accuracy metrics. All models demonstrated systematic demographic biases: married individuals were consistently under-predicted across all LLMs, with 9 statistically significant biases detected after Bonferroni correction. We introduce RFI as a comprehensive evaluation framework decomposing model performance into magnitude accuracy, direction consistency, pattern preservation, and scale calibration. These findings illuminate critical blind spots in current evaluation practices and provide an actionable framework for bias-aware assessment of LLMs in social research applications.

1 Introduction

Large Language Models (LLM) increasingly serve as synthetic participants in social science research. Recent work suggests that LLM-generated synthetic data can be biased that standard evaluation approaches may not easily capture. For instance, Argyle et al. (2023) demonstrate that LLM outputs reflect real-world biases, while Wang et al. (2025) reveal "group flattening"—oversimplified representation of demographic groups that diminishes within-group heterogeneity. Hu et al. (2024) show that models inherently exhibit social identity biases, potentially reinforcing existing societal inequalities, and Shrestha et al. (2025) find that while LLMs capture broad aggregate trends, they systematically fail to replicate the nuance and precision characteristic of authentic human responses.

Current evaluation approaches focus on aggregate accuracy metrics, but for the social and political science, more sophisticated evaluations are needed. This paper proposes a metric that is made up of four components: a) magnitude accuracy, b) directional consistency of estimated coefficients, c) pattern preservation, and d) systematic scale bias. In order to demonstrate the usefulness of this evaluation metric, we focus in on a particular methodological approach commonly used in the social and political sciences: Factorial Experiments. This method is extremely important in political science, because it estimates nuanced attitude and behavior patterns. We use a particular study of family ideals vignettes to demonstrate the usefulness of the composite evaluation metric. Aassve et al. (2024) surveyed 20,141 respondents across eight countries. We use the resulting data from this survey to evaluate the extent LLMs can authentically replicate established human preference patterns when defined over a multi-dimensional concept. This provides a rigorous methodological foundation: we leverage validated cross-cultural factorial survey experiments (Auspurg and Hinz, 2015) that systematically vary family characteristics across ten dimensions (i.e. factors) to assess genuine replication fidelity.

Our methodological innovation centers on experimental fidelity rather than prompt optimization. We preserve the exact vignette text and rating instructions presented to human respondents, adding only demographic persona information derived from authentic

participant profiles. This approach eliminates prompt engineering confounds and provides unbiased assessment of model suitability for real research applications—testing whether LLMs can handle authentic research protocols rather than engineered tasks.

We employ five state-of-the-art LLMs (GPT-4, GPT-4.1, DeepSeek EN/CN, Claude 3.5) across two very different cultural contexts, China and the US. Our contributions are threefold: (1) We show that successful replication requires capturing human variability and disagreement, not just central tendencies; (2) We introduce the Replication Fidelity Index as a comprehensive bias-aware evaluation framework that decomposes model performance into interpretable components; (3) We demonstrate systematic demographic biases across all models, with married individuals consistently under-predicted, revealing critical blind spots invisible to standard accuracy metrics.

2 Methodology

Experimental Design and Data. We analyze vignette experiments from China (n=5,186) and the United States (n=5,906) following Aassve et al. (2024). Respondents evaluated systematically varied family scenarios on a 0–10 scale of "successful family." Eight attributes were manipulated: (1) union status (married/cohabiting), (2) number of children (0–3+), (3) income (below/average/above), (4) family communication (poor/good), (5) grandparent contact (infrequent/frequent), (6) community respect (not respected/respected), (7) gender roles (traditional/commonplace/egalitarian), and (8) work–family conflict (male/female/neither/both conflicted). We estimated Average Component Marginal Effects (ACMEs) via fixed-effects models.

Experimental Fidelity Protocol. LLMs received the exact vignette text and rating instructions as human respondents ("On a scale of 0–10, to what extent does this describe a successful family?"). To simulate realistic heterogeneity, we incorporated demographic personas directly from authentic survey profiles (e.g., age, education, income, family structure). This ensures evaluation reflects real-world respondent distributions rather than engineered prompts (Full prompt and examples of persona and vignette in Appendix C).

Replication Fidelity Index (RFI). To compare model and human responses, we propose the Replication Fidelity Index, which decomposes performance into four bounded components: (A) magnitude accuracy (RMSE of normalized effects), (B) directional consistency (share of signs correctly predicted), (C) pattern preservation (correlation of relative factor importance), and (D) scale bias (systematic inflation/deflation). These components are combined into an interpretable index ranging from 0 (failure) to 1 (perfect replication). Formal definitions and proofs of boundedness appear in Appendix B.

The final RFI score combines these components: $\text{RFI} = 1 - [w_A \cdot A + w_B \cdot B + w_C \cdot C + w_D \cdot D]$ where all components are bounded in [0,1], weights sum to 1, and the index provides interpretable scores from 0 (complete failure) to 1 (perfect replication). Mathematical details and formal boundedness proofs are in Appendix B.

Bias Detection Framework. To identify systematic demographic biases, we complemented vignette analysis with subgroup tests, applying Bonferroni-corrected Fisher's exact tests across demographic categories using Fisher's exact tests with Bonferroni correction across 567 statistical comparisons. The 567 tests represent all viable combinations after filtering: $(5 \text{ models} \times \text{demographic categories} \times 2 \text{ bias directions})_{\text{China}} + (5 \text{ models} \times \text{demographic categories} \times 2 \text{ bias directions})_{\text{USA}}$, excluding combinations with insufficient data.

3 Results

Critical Variability Gap Reveals Fundamental Limitation. Our analysis reveals a counterintuitive finding that challenges standard evaluation approaches: successful replication requires capturing human disagreement patterns, not just central tendencies. Table 1 demonstrates this fundamental limitation—human ratings exhibit natural variance (σ = 2.4-2.6) with full 0-10 scale usage, while all LLMs show severely constrained distributions.

DeepSeek models are particularly constrained (σ = 1.1-1.3, limited to 3-8 range), while GPT-4 comes closest to human variability (σ = 2.0) but still falls short of authentic disagreement patterns.

Table 1: Summary Statistics Reveal Critical Variability Gap

Model	China (n=5,186)			USA (n=5,906)		
	Mean	Std	Range	Mean	Std	Range
Ground Truth	5.27	**2.40**	0-10	5.88	**2.64**	0-10
DeepSeek (CN)	5.63	**1.10**	3-8	5.63	**1.15**	3-9
DeepSeek (EN)	5.45	**1.17**	3-9	5.37	**1.29**	3-9
GPT-4	5.52	**2.02**	2-9	5.52	**2.07**	2-10
Claude 3.5	4.48	**1.34**	2-9	5.00	**1.56**	2-10

This variability constraint represents a systematic failure across all tested models—even those with accurate mean predictions miss the natural human response patterns essential for valid social science applications. Models appear to be overly confident, clustering ratings in narrow ranges rather than reflecting authentic human disagreement about complex social judgments.

Table 2: RFI Component Analysis Reveals Model-Specific Strengths and Vulnerabilities

Model	**A (Magnitude)**	**B (Direction)**	**C (Pattern)**	**D (Scale)**	**RFI Score**
USA Results					
DeepSeek CN	0.167	0.143	**0.044**	0.172	**0.869**
DeepSeek EN	0.135	0.143	0.071	0.174	**0.869**
GPT-4	0.121	**0.000**	0.131	0.318	0.858
Claude 3.5	0.135	0.143	0.111	0.239	0.843
China Results					
DeepSeek CN	0.147	0.214	0.132	**0.062**	**0.861**
Claude 3.5	0.158	0.143	0.226	0.106	0.842
DeepSeek EN	0.135	0.286	0.224	0.043	0.828
GPT-4.1	0.131	0.143	0.213	0.240	0.818

Note: Component A measures magnitude accuracy (weighted RMSE), B measures directional accuracy (proportion of sign errors), C measures pattern accuracy (1-—correlation—), D measures systematic scale bias. Lower component scores indicate better performance; higher RFI scores indicate superior replication fidelity.

RFI Reveals Component-Specific Model Limitations. Table 2 shows comprehensive model performance across our four-component framework, revealing systematic patterns invisible to aggregate metrics. DeepSeek models achieve highest overall RFI scores (0.869 USA, 0.861 China) through superior pattern preservation and scale calibration, despite their constrained variability. This apparent paradox—high RFI with low variability—demonstrates that RFI rewards *replication fidelity of preference structures* rather than variability matching. DeepSeek correctly identifies directional relationships and preserves relative factor importance even within constrained ranges, while models with higher variability may miss systematic preference patterns. Notably, models excel in different components—GPT-4 achieves perfect directional accuracy in the USA (B=0.000) but suffers from severe scale bias (D=0.318), while DeepSeek models demonstrate superior pattern preservation. This component-wise analysis reveals that model selection should depend on research priorities: directional accuracy may matter more than magnitude precision for some applications.

Systematic Demographic Biases Across All Models. Our rigorous bias detection framework revealed 9 statistically significant demographic biases after Bonferroni correction across 567 comprehensive statistical tests (Table 3). Union status emerges as the dominant source of systematic bias (7 of 9 findings), with married individuals consistently under-predicted across all models in the USA, while single individuals suffer under-prediction by DeepSeek models in China. Additionally, Claude 3.5 exhibits pronounced gender-specific biases in the USA, systematically under-predicting both male and female ratings through different mechanisms.

Table 3: Systematic Demographic Biases Detected Across All Models

Country	Model	Demographics	Bias Direction	P-value
China	DeepSeek CN	Union: single	Under-predicted	0.005
China	DeepSeek EN	Union: single	Under-predicted	0.032
USA	Claude 3.5	Gender: female/male	Under-predicted	0.002
USA	Claude 3.5	Union: married	Under-predicted	<0.001
USA	DeepSeek CN	Union: married	Under-predicted	<0.001
USA	DeepSeek EN	Union: married	Under-predicted	<0.001
USA	GPT-4.1	Union: married	Under-predicted	<0.001
USA	GPT-4	Union: married	Under-predicted	<0.001

Critically, all identified biases involve under-prediction rather than over-prediction, suggesting systematic underestimation patterns that would compromise research conclusions. These biases demonstrate that models have consistent "blind spots" for specific population groups—patterns that would be completely invisible to standard accuracy-based evaluation approaches yet fundamentally undermine research validity when LLMs serve as synthetic participants.

4 Discussion and Implications

The counter-intuitive insight that successful replication requires capturing human disagreement fundamentally challenges standard evaluation approaches focused on central tendencies. This variability gap represents a systematic failure across all tested models—even those with accurate mean predictions miss the natural human response patterns essential for valid social science applications.

The systematic demographic biases demonstrate that all models have consistent blind spots for specific population groups. The pervasive under-prediction of married individuals across all models suggests either training data artifacts or systematic processing limitations that compromise research validity. Importantly, these biases would be completely missed by standard accuracy metrics that focus on aggregate performance rather than demographic fairness, highlighting the urgent need for bias-aware evaluation frameworks in AI-assisted social research.

Our RFI framework provides actionable methodology for detecting these hidden systematic limitations. The component-wise analysis reveals that models fail in fundamentally different ways: GPT-4 achieves perfect directional understanding but suffers from scale bias, while DeepSeek models preserve relative importance patterns despite constrained variability. This suggests that model selection should depend on specific research contexts and priorities rather than aggregate performance scores.

Actionable Recommendations. (1) Use multi-component frameworks like RFI to assess replication beyond aggregate accuracy; (2) Test systematically for demographic bias with proper statistical corrections before deployment; (3) Prioritize experimental fidelity over prompt tuning to evaluate genuine model capability; (4) Match model choice to research priorities—e.g., directional accuracy for exploration, pattern preservation for policy.

5 Limitations

Our study has several constraints. First, we analyze only two national contexts (China and the US). While culturally distinct, broader cross-national testing is needed to assess generalizability. Second, we restrict analysis to family ideals; other domains (e.g., political attitudes, health behaviors) may present different replication challenges. Third, we test a subset of current LLMs; results may shift with new releases, though the systematic patterns observed suggest deeper structural issues. Finally, our vignette-based design cannot capture all aspects of real survey interaction, such as interviewer effects or open-ended responses.

References

A. Aassve, A. Adserà, P. Y. Chang, L. Mencarini, H. Park, C. Peng, S. Plach, J. M. Raymo, S. Wang, and W.-J. J. Yeung. Family ideals in an era of low fertility. *Proceedings of the National Academy of Sciences*, 121(6):e2311847121, 2024. doi: 10.1073/pnas.2311847121.

A. Agresti. *An Introduction to Categorical Data Analysis*. John Wiley & Sons, 3rd edition, 2018.

L. P. Argyle, E. C. Busby, N. Fulda, J. Gubler, C. Rytting, and D. Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3): 317–334, 2023. Preprint available at arXiv:2209.06899, 2022.

K. Auspurg and T. Hinz. *Factorial Survey Experiments*. Quantitative Applications in the Social Sciences. SAGE Publications, Thousand Oaks, CA, 2015.

T. Hu, N. Levy, M. Lazar, S. Harrell, K. Vafa, C. A. Hidalgo, and H. Shirado. Generative language models exhibit social identity biases. *Nature Computational Science*, 4(11): 741–751, 2024. doi: 10.1038/s43588-024-00741-1.

P. Shrestha, D. Krpan, F. Koaik, R. Schnider, D. Sayess, and M. S. Binbaz. Beyond weird: Can synthetic survey participants substitute for humans in global policy research? *Behavioral Science & Policy*, 2025.

A. Wang, J. Morgenstern, and J. P. Dickerson. Large language models as replacements for human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 2025. doi: 10.1038/s42256-025-00986-z.

A Data Quality Checks (Grouped persona data - 6 vignettes / persona)

Table 4: Distributional Validation Summary for Persona-Level Ratings

	China (n=866)			USA (n=986)		
Rating Source	**Range**	**D'Agostino p**	**Outliers**	**Range**	**D'Agostino p**	**Outliers**
Ground Truth	[0.83, 9.83]	0.231	4	[0.83, 9.83]	0.398	7
Claude 3.5	[3.67, 5.50]	0.986	22	[3.83, 5.83]	<0.001	15
DeepSeek CN	[5.00, 6.20]	0.003	3	[5.00, 6.40]	0.471	1
DeepSeek EN	[4.75, 5.83]	0.158	37	[4.80, 5.83]	0.035	71
GPT-4.1	[4.00, 5.83]	<0.001	11	[4.17, 6.20]	<0.001	20
GPT-4	[4.25, 6.33]	<0.001	12	[4.50, 6.33]	<0.001	17

Notes: D'Agostino test assesses normality ($p > 0.05$ supports normality). Outliers detected using IQR method (beyond 1.5×IQR from Q1/Q3). Ground truth shows appropriate spread and minimal outliers. DeepSeek models exhibit constrained ranges, consistent with distributional analysis in Section 2.

B Mathematical Framework

Table 5: RFI Component Framework

Component	**Mathematical Definition**
Magnitude (A)	$\sqrt{\frac{\sum_{k=1}^{K} w_k(\tilde{b}_k^{\text{LLM}} - \tilde{b}_k^{\text{GT}})^2}{4}}$
Direction (B)	$\frac{1}{K}\sum_{k=1}^{K} \mathbf{1}[\text{sign}(t_k^{\text{LLM}}) \neq \text{sign}(t_k^{\text{GT}})]$
Pattern (C)	$1 - \lvert\text{corr}(\tilde{\mathbf{b}}^{\text{LLM}}, \tilde{\mathbf{b}}^{\text{GT}})\rvert$
Scale Bias (D)	$\frac{\lvert R-1\rvert}{R+1}$ where $R = \frac{\sum_{k=1}^{K} \lvert b_k^{\text{LLM}}\rvert}{\sum_{k=1}^{K} \lvert b_k^{\text{GT}}\rvert}$

B.1 Formal RFI Component Definitions

Let $\mathbf{b}^{\text{GT}} = (b_1^{\text{GT}}, b_2^{\text{GT}}, \ldots, b_K^{\text{GT}})$ be our ground truth coefficient vector from human survey data with standard errors $\boldsymbol{\sigma}^{\text{GT}} = (\sigma_1^{\text{GT}}, \sigma_2^{\text{GT}}, \ldots, \sigma_K^{\text{GT}})$ and $\mathbf{b}^{\text{LLM}} = (b_1^{\text{LLM}}, b_2^{\text{LLM}}, \ldots, b_K^{\text{LLM}})$ be the corresponding LLM-generated coefficients with standard errors $\boldsymbol{\sigma}^{\text{LLM}}$.

T-Statistic Calculation. Statistical significance weighting ensures that factors with strong, reliable human effects receive appropriate emphasis in evaluation:

$$t_k^{\text{GT}} = \frac{b_k^{\text{GT}}}{\sigma_k^{\text{GT}}}, \qquad t_k^{\text{LLM}} = \frac{b_k^{\text{LLM}}}{\sigma_k^{\text{LLM}}}$$

Joint Coefficient Normalization. To enable meaningful comparisons across different effect scales:

$$\tilde{b}_k^{\text{GT}} = \frac{b_k^{\text{GT}}}{S}, \qquad \tilde{b}_k^{\text{LLM}} = \frac{b_k^{\text{LLM}}}{S}$$

where $S = \max\{\max_i |b_i^{\text{GT}}|, \max_i |b_i^{\text{LLM}}|\}$ ensures all normalized coefficients lie in $[-1, 1]$.

Importance Weights. T-statistic based weighting prioritizes factors where humans show clear, consistent patterns:

$$w_k = \frac{|t_k^{\text{GT}}|}{\sum_{i=1}^{K} |t_i^{\text{GT}}|}$$

B.2 RFI Framework Design and Boundedness

The RFI is designed to provide interpretable evaluation scores between 0 (complete failure) and 1 (perfect replication). Each component is carefully bounded:

Magnitude Component (A): Measures how wrong effect sizes are using weighted RMSE of normalized coefficients. Since coefficients are normalized to $[-1, 1]$, maximum possible difference is 2, and we divide by 4 to ensure $A \in [0, 1]$.

Direction Component (B): Counts proportion of factors where LLM gets basic positive/negative relationship wrong. As a proportion, B is naturally bounded in $[0, 1]$.

Pattern Component (C): Uses correlation transformation $C = 1 - |\rho|$ to measure whether LLM preserves relative factor importance. Since correlation coefficients satisfy $|\rho| \leq 1$, we have $C \in [0, 1]$.

Scale Bias Component (D): Detects systematic inflation/deflation using $D = \frac{|R-1|}{R+1}$ where R is ratio of total effect sizes. This function maps all positive values to $[0, 1)$.

The final RFI combines these bounded components with weights summing to 1, ensuring RFI $\in [0, 1]$ with intuitive interpretation: higher scores indicate better replication.

B.3 Statistical Methodology

Bootstrap Validation. We conducted parametric bootstrap analysis (2,000 iterations) sampling coefficients from normal distributions using original standard errors, then recalculating complete RFI pipelines. Results show that while apparent ranking differences exist, only subset are statistically significant: DeepSeek models significantly outperform GPT-4.1 in USA (p=0.035-0.038) and GPT-4 models in China (p¡0.027), with performance gaps of 0.057-0.090 RFI points.

Weight Sensitivity Analysis. Model rankings depend on theoretical priorities, but core conclusions remain robust. DeepSeek models demonstrate consistent performance across different weighting schemes, while GPT-4 exhibits extreme sensitivity (best to worst performer depending on priorities). This validates that conclusions aren't artifacts of arbitrary weight choices while highlighting the importance of component-specific evaluation for practical model selection.

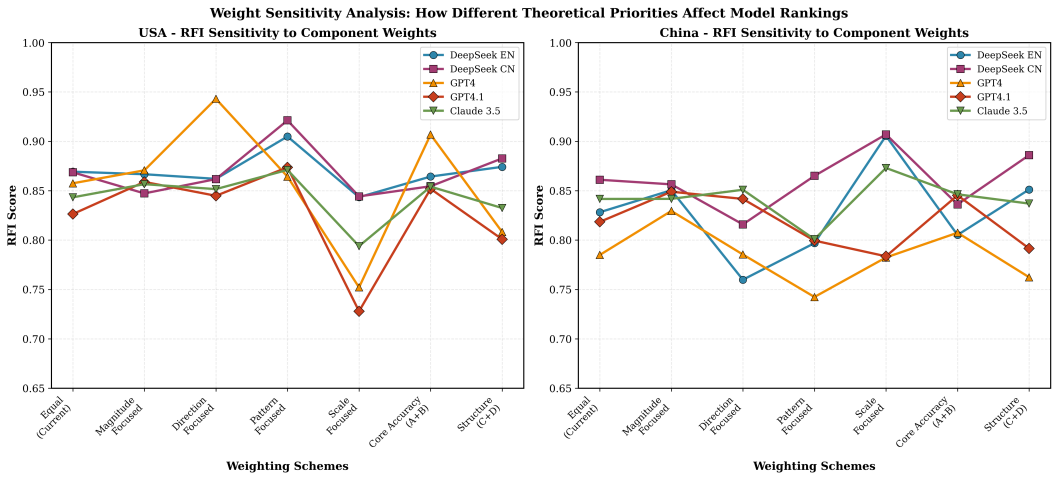

Figure 1: Weight sensitivity

C Experimental Details

Model Specifications. We evaluated five state-of-the-art models: GPT-4 (March 2023, 8K-32K context), GPT-4.1 (April 2025, 1M context), DeepSeek V3 EN/CN (December 2024, 128K context), and Claude 3.5 Sonnet (June 2024, 200K context). Models were selected to cover capability spectrum, cultural alignment, and cost considerations.

Persona Construction. Demographic personas were derived directly from authentic participant profiles in the original PNAS dataset, maintaining realistic combinations of demographic characteristics. Each persona includes age, education level, income, family structure, employment status, and cultural context drawn from actual survey respondents.

Statistical Testing Framework. Demographic bias detection employed Fisher's exact tests across 567 comparisons with Bonferroni correction (p-values multiplied by number of tests). We used 2×2 contingency tables (Agresti, 2018) comparing each demographic group's representation in bias categories versus background population, with odds ratios calculated as OR = ad/bc.

Sample prompt

Listing 1: User Prompt Configuration

```
user_prompt = (
 "**Task**: You will be presented with six descriptions of different families. "
"Please answer the questions following the description of each family. "
"Please take your time picturing each family while reading the description.\n"
f"{vignette}\n"
"**Question**: Based on this description, on a scale of 0 (0 = Strongly Disagree) to 10 (10 = Strongly Agree) "
"to what extent do you agree or disagree with the following statement? "
'"This describes a successful family." '
"Combining the description and your own characteristics, please respond with a single number between 0-10, no explanation just number.\n**Answer**: "
)
```

Examples of Persona and Vignette

PERSONA:
"You are a respondent in the LifePoints platform who answers surveys for money. You are currently answering a comparative survey about family ideals that is conducted in 8 countries including the USA, Italy, Spain, Norway, urban China, Singapore, Japan, and Korea. You are a participant located in China. You have the following characteristics.. You are a 41.0 year-old male. currently living in big city and its suburbs. Your race is 999. and your ethnicity is 999. You do not belong to any religious denomination. You are currently married. You have been in a consensual union before. and have been married for 2003 years. You cohabited with your spouse before marriage. Your highest level of education is ISCED level 6. Your current work situation is: Employed. working as Professional and technical. with a Fixed-term contract. You have never left your parents' home. Your household consists of 3 people. You have 1 brother. and are the firstborn. You have 2.0 children. and definitely no plan to have children in the next 3 years. You have no gender preference if you could have only one child. Your monthly household income after taxes is equivalent to 2,800.00 euros. You spent 48.30 seconds in reading, judging and rating this description.. On a scale from 0 to 10, you rated satisfaction with your life a 8."

VIGNETTE:
"This is the first family description. In the following you will find a description of Lisa and Robert. Lisa and Robert are both around 45 years old. Lisa and Robert are married. Lisa

and Robert have three children. Lisa and Robert's combined income is higher than the country average. The family is not well respected in their community. Lisa, Robert, and their children discuss their daily life frequently and feel comfortable expressing their feelings and raising disagreements with each other. Lisa, Robert, and their children talk with both Lisa's and Robert's parents infrequently. Both Lisa and Robert work full-time. Lisa takes care of most of the family and household responsibilities. Robert feels conflicted between his career and the possibility to help out with family responsibilities, and Lisa also feels conflicted between her family responsibilities and her career."