Predictive Completeness of Social-Preference Theories

Jesper Armouti-Hansen*

June 21, 2026

Abstract

[Placeholder — written last, against the final results. The paper finalizes Armouti-Hansen (2022): we measure the completeness of a nested ladder of social-preference theories on the data of Bruhin et al. (2019), at two information sets — the representative-agent level and, conditioning all three legs on subject identity, the individual level — using a held-out machine-learning frontier to bound the unrecoverable individual-level oracle.]

*Email: jesper@armoutihansen.xyz.

1 Setup

This section fixes the primitives, following Fudenberg et al. (2022), and defines a model's *completeness* at the two information sets the paper evaluates. We then describe the data, introduce the social-preference theories, and translate each into a probabilistic forecast.

1.1 Primitives and completeness

A population is described by the joint distribution P of a triple (Y, X, N). The outcome $Y \in \{0, 1\}$ indicates which of two allocations a subject selects. The choice situation X takes values in a finite set $\mathcal{X} \subset \mathbb{R}^d$ and collects the attributes of the two alternatives on offer; its components are made explicit in Section 1.4. The identifier $N \in \mathcal{N}$, with $\mathcal{N}$ finite and $P(N = n) > 0$ throughout, names the subject who faces the situation. We assume the conditional laws below are non-degenerate.

A forecast maps an information set to a probability that $Y = 1$ and is scored by its log loss,

$$\ell(y, p) = -\big[\, y \log p + (1 - y) \log(1 - p) \,\big], \tag{1}$$

with $E_f := \mathbb{E}_P[\ell(Y, f)]$ the risk of a forecast f. Minimising expected log loss coincides with minimising the expected Kullback–Leibler divergence from the true conditional law, so risk is minimised by the true conditional probability of Y given the conditioning information. We call that forecast the *oracle* and its risk the *irreducible risk* E^*. The risk of any forecast then splits into the irreducible risk and a non-negative excess,

$$E_f = E^* + d(f), \qquad d(f) \geq 0, \tag{2}$$

where $d(f)$ is f's mean divergence from the oracle. In finite samples $d(f)$ reflects two distinct shortfalls: a parametric forecast may be unable to represent the oracle however much data it sees (bias), and a flexible forecast fit on limited data may track sampling noise (variance). The second shortfall also makes the irreducible risk itself hard to recover, since a flexible estimate of the oracle can overfit; an estimated oracle risk is then an upper bound on the truth rather than the truth. Section 2 returns to this where it binds.

Fudenberg et al. (2022) locate a model between two reference forecasts. The lower reference is a *naive baseline*, the best forecast that uses none of the situational information; the upper reference is the oracle. A model's completeness is the fraction of the baseline-to-oracle risk gap it closes, so it is zero when the model merely matches the baseline, one when it matches the oracle, and well defined whenever the situation is predictive (the oracle strictly beats the baseline). We apply the measure under a single discipline: the baseline, the model, and the oracle that enter a completeness ratio all live on the same information set. That information set is the question being asked, and we ask it at two levels.

Definition 1.1 (Representative-agent completeness)**.** On information set X, with $H(Y) > H(Y \mid X)$, the representative-agent completeness of a forecast f is

$$C_{\mathrm{RA}}(f) = \frac{H(Y) - E_f(X)}{H(Y) - H(Y \mid X)}, \tag{3}$$

the baseline $H(Y)$ being the unconditional choice rate and the oracle $H(Y \mid X)$ the situation-conditional irreducible risk.

Definition 1.2 (Individual-level completeness)**.** On information set (X, N), with $H(Y) > H(Y \mid X, N)$, the individual-level completeness of a forecast f is

$$C_{\mathrm{IL}}(f) = \frac{H(Y) - E_f(X, N)}{H(Y) - H(Y \mid X, N)}, \tag{4}$$

the baseline $H(Y)$ being the same no-information choice rate as at the representative-agent level and the oracle $H(Y \mid X, N)$ the individual-level irreducible risk.

Representative-agent completeness is the measure of Fudenberg et al. (2022) as ordinarily applied: it reports the fraction of the predictable variation in *aggregate* choice rates that a theory captures. Individual-level completeness applies the *same* measure to the richer information set (X, N). The baseline is unchanged — the no-information rate $H(Y)$ — and only the oracle moves, from the situation oracle $H(Y \mid X)$ to the individual-level oracle $H(Y \mid X, N)$. It reports the fraction of the predictable variation in *each subject's* choices, situation- and subject-driven together, that the theory captures: the level at which a theory of individual behaviour is meant to hold. The two measures share a baseline and differ only in the information their oracle may use, and they coincide when subjects are exchangeable (N independent of (X, Y)).

A decomposition ties the two together and makes the case for the second. Write $I(Y; X) = H(Y) - H(Y \mid X)$ and $I(Y; X, N) = H(Y) - H(Y \mid X, N)$ for the predictable variation available at the two levels, with $I(Y; X) \leq I(Y; X, N)$. For the representative-agent rung f_1 — a theory evaluated as a single common parameter, which ignores N and so carries the same risk on both information sets — the two scores are related exactly by

$$C_{\mathrm{IL}}(f_1) = \frac{I(Y; X)}{I(Y; X, N)} C_{\mathrm{RA}}(f_1). \tag{5}$$

The representative-agent score is rescaled by $I(Y; X)/I(Y; X, N)$, the share of the individual predictable variation that is situation-driven. The factor is at most one, and strictly below one whenever subject identity carries predictive information beyond the situation; the representative-agent verdict then *overstates* how complete the theory is as an account of individuals, by exactly the share of predictable variation it cannot reach.

The same decomposition reads the heterogeneity ladder. A model f that lets the parameter vary across subjects — a K-type or mixed-logit version of the theory whose representative-agent rung is f_1 — has individual-level completeness

$$C_{\mathrm{IL}}(f) = \underbrace{\frac{I(Y;X)}{I(Y;X,N)}\, C_{\mathrm{RA}}(f_1)}_{\text{structural}} + \underbrace{\frac{E_{f_1}(X) - E_f(X,N)}{I(Y;X,N)}}_{\text{heterogeneity fit}}, \tag{6}$$

the theory's representative-agent completeness, rescaled as above, plus the non-negative gain from fitting individual heterogeneity.[1] The ladder $f_1 \to f_K \to f_G$ thus climbs the individual-level scale entirely through the second term: every rung shares the structural floor $[I(Y;X)/I(Y;X,N)]\, C_{\mathrm{RA}}(f_1)$ and is separated only by how much individual heterogeneity it captures.

This is why the representative-agent verdict cannot stand in for the individual one. Social-preference theories are theories of *individual* choice: each holds that a person evaluates allocations through a utility carrying her own parameters. A theory can reproduce the average choice rate in every situation, earning a high C_{RA}, while predicting no individual well; by (5) the aggregate score is then inflated by precisely the individual predictable variation the situation average omits. To report only C_{RA} is to answer a question the theory does not pose. The individual-level measure evaluates the theory where it operates, and the two agree only when there is no such variation to omit.

The levels differ sharply in how their oracle can be estimated. The situation oracle $H(Y \mid X)$ is recoverable: with many subjects facing each situation, the situation-conditional choice rate is well estimated (Section 2.2). The individual-level oracle $H(Y \mid X, N)$ is not. In the panel of Section 1.2 every (subject, situation) cell is observed once, so the per-cell choice rate is degenerate and no consistent estimate of $H(Y \mid X, N)$ exists without the within-cell replication the design rules out. We therefore do not estimate the individual-level oracle but bound it, replacing it by the best held-out risk $\bar{E}$ of a set of flexible pooling predictors, the ML frontier of Section 2.3. The next remark records what that substitution secures.

Remark 1.1 (Frontier bound). The oracle $H(Y \mid X, N)$ is the lowest risk any forecast on

[1] By Definition 1.1, $H(Y) - E_{f_1}(X) = C_{\mathrm{RA}}(f_1)\, I(Y;X)$; writing $H(Y) - E_f(X,N) = \big(H(Y) - E_{f_1}(X)\big) + \big(E_{f_1}(X) - E_f(X,N)\big)$ and dividing by $I(Y;X,N)$ gives (6). The rung f_1 ignores N, so $E_{f_1}(X,N) = E_{f_1}(X)$, the second term vanishes, and (6) reduces to (5).

(X, N) can attain (Gibbs' inequality). Unable to recover it, we substitute the lowest risk we can *construct*: the frontier $\bar{E}$, the best held-out risk among the flexible predictors of Section 2.3, which by the same inequality satisfies $\bar{E} \geq H(Y \mid X, N)$. Replacing the oracle by $\bar{E}$ shrinks the denominator of (4), so for every forecast f

$$\widetilde{C}_{\mathrm{IL}}(f) = \frac{H(Y) - E_f(X, N)}{H(Y) - \bar{E}} \;\geq\; C_{\mathrm{IL}}(f),$$

making a reported individual-level completeness an upper bound on the truth and a reported shortfall a lower bound on the true shortfall. Two qualifications matter. First, the inequality is about *population* risks, and Gibbs' inequality binds the structural forecasts no less than the frontier; what makes the frontier the right denominator is not the inequality but that its risk is the lowest we can achieve, the tightest available stand-in for the oracle. Second, the operative $\bar{E}$ is an *estimate*, so the bound transfers to it only in expectation: we guard against selection optimism by choosing the frontier on a validation slice and scoring it on a separate held-out fold, never on the rows whose risk is reported (Section 2.3), and a $\widetilde{C}_{\mathrm{IL}}$ above one — a structural rung beating the estimated frontier — is reported, not clipped. Because the same $\bar{E}$ enters every denominator, the ranking of forecasts and the comparisons across ladder rungs do not depend on where the ceiling lies.

Individual-level completeness can be read across subjects as well as in aggregate. Once a model sorts subjects into types (Section 1.4), the measure restricted to the members of a given type — the same numerator and the same frontier ceiling, evaluated on those subjects — is a *within-type completeness*, and it localises where a theory's residual incompleteness sits: which preference types the theory captures and which it leaves predictable variation in. We treat it as a diagnostic decomposition of C_{IL} rather than a headline measure, and give its construction in Section 2, where types are estimated.

We use no further formal apparatus; the chain rule $I(Y; X, N) = I(Y; N) + I(Y; X \mid N)$, which links the two levels' denominators, serves only as interpretation.

1.2 Data

We draw on the laboratory experiment of Bruhin et al. (2019). Subjects, students at the University of Zürich, took part in two sessions three months apart and faced the same menu of binary allocation problems in each, always in the active role of Player A. Our main panel is the first session: 174 subjects, each making 117 choices, for 20,358 observations. The second session repeats the design and serves as a replication arm. The experiment also recorded demographic and personality measures (age, gender, cognitive ability, and the Big Five), which we use to describe the estimated preference types but not to forecast choices.[2]

Of the 117 problems, 39 are dictator problems, in which Player A chooses between two allocations a and b, each specifying a payoff π^A for herself and π^B for a passive Player B. One third of these problems keep A ahead of B under either allocation ($\pi^A > \pi^B$), one third keep A behind ($\pi^A < \pi^B$), and one third trade A's payoff off against B's. Within each block the cost to A of altering B's payoff is varied so that the inequity-aversion (differentiated-altruism) parameters are identified over the range $[-3, 1]$.

The remaining 78 problems are reciprocity problems, elicited by the strategy method. Player B first chooses whether to implement a reference allocation $c = (\pi_c^A, \pi_c^B)$; if she does not, Player A chooses between a and b, which reuse the dictator allocations. Half the problems are *negative*, meaning that B's declining c leaves A strictly worse off under both a and b, so the move reads as unkind; half are *positive*, leaving A strictly better off, so the move reads as kind. Problems in which A would be better off under one of a, b and worse off under the other are not used.

The source study's analysis sample drops 14 subjects whose choices show near-zero sensitivity to payoffs, leaving 160. We report the full 174-subject panel in the main tables and the 160-subject sample as a sample-alignment robustness check. Because each (subject, situation) cell is observed once, the individual-level oracle is unrecoverable and the bound of Remark 1.1 is the operative reporting device throughout.

1.3 Social preference models

We evaluate a nested ladder of social-preference theories, each obtained from the previous one by admitting a further other-regarding motive. A *theory* specifies Player A's systematic utility u from an allocation as a function of the two payoffs and the context; it is silent on how choice becomes stochastic and on how subjects differ, both of which are added in Section 1.4. Let $\mathbf{1}_B$ and $\mathbf{1}_A$ indicate that A is *behind* ($\pi^B > \pi^A$) and *ahead* ($\pi^B < \pi^A$), and let $\mathbf{1}_K$ and $\mathbf{1}_U$ indicate that B's preceding move in a reciprocity problem was kind or unkind.

The ladder begins with pure self-interest,

$$u^{\text{self}} = \pi^A, \tag{7}$$

which carries no parameters. It is the ladder's first rung, not the completeness baseline: the baseline is the no-situation optimum of Section 1.1. Simple altruism lets B's payoff enter A's

[2] Choices were incentivised: each subject's payment combined a show-up fee, a fixed amount for the questionnaire, and the realised outcomes of three randomly drawn decisions, averaging 52.5 CHF in the first session (Bruhin et al., 2019).

utility with a single weight, whether A is ahead or behind,

$$u^{\mathrm{alt}} = \pi^A + \gamma_S\,(\pi^B - \pi^A), \tag{8}$$

where $\gamma_S > 0$ is altruism and $\gamma_S < 0$ is spite.[3]

Differentiated altruism, a variant of the inequity-aversion model of Fehr and Schmidt (1999), allows that weight to differ according to whether A is behind or ahead,

$$u^{\mathrm{diff}} = \pi^A + (\gamma_D\,\mathbf{1}_B + \gamma_A\,\mathbf{1}_A)(\pi^B - \pi^A). \tag{9}$$

We leave γ_D and γ_A unrestricted. A negative γ_D is aversion to being behind (or spite when behind) and a positive γ_D is altruism when behind; a positive γ_A is aversion to being ahead (altruism when ahead) and a negative γ_A is spite when ahead.[4]

The last two rungs add reciprocity, which the design encodes through the binary kind/unkind context. Following Charness and Rabin (2002), an unkind preceding move shifts the inequity-aversion weights,

$$u^{\mathrm{cr}} = \pi^A + (\gamma_D\,\mathbf{1}_B + \gamma_A\,\mathbf{1}_A + \gamma_U\,\mathbf{1}_U)(\pi^B - \pi^A), \tag{10}$$

so that $\gamma_U < 0$ is negative reciprocity. The full specification of Bruhin et al. (2019) lets a kind move shift them too,

$$u^{\mathrm{full}} = \pi^A + (\gamma_D\,\mathbf{1}_B + \gamma_A\,\mathbf{1}_A + \gamma_K\,\mathbf{1}_K + \gamma_U\,\mathbf{1}_U)(\pi^B - \pi^A), \tag{11}$$

adding $\gamma_K > 0$ as positive reciprocity. The five theories nest in sequence: each utility specialises to the previous one when its new parameters are set to zero.[5]

1.4 Parametric models

A theory becomes a *model*, and so a forecast, once we add a stochastic choice rule and a heterogeneity structure. We first make the situation concrete. A problem g presents two

[3] Writing the other-regarding term in $(\pi^B - \pi^A)$ rather than π^B leaves the forecast unchanged but makes γ_S directly comparable to the unit weight on A's own payoff.

[4] Leaving the signs free matters on this data: imposing Fehr–Schmidt behindness aversion drives its estimate toward zero at a clear cost in fit, which is also why we do not adopt the Bolton and Ockenfels (2000) specification.

[5] We also estimated the reciprocal-altruism specification of Levine (1998), which adds reciprocity to *simple* altruism, $(\gamma_S + \gamma_K\mathbf{1}_K + \gamma_U\mathbf{1}_U)(\pi^B - \pi^A)$. Lacking the behind/ahead distinction, it does not nest within the ladder, and it predicts less well than differentiated altruism; we report it only as a robustness comparison.

allocations, and allocation $m \in \{a, b\}$ is described by the feature vector

$$x_{g,m} = \big(\pi^A_{g,m},\, \pi^B_{g,m},\, \mathbf{1}_{B,g,m},\, \mathbf{1}_{A,g,m},\, \mathbf{1}_{K,g},\, \mathbf{1}_{U,g}\big),$$

so the situation is the pair $X = (x_{g,a}, x_{g,b})$.[6]

For the choice rule we follow Bruhin et al. (2019) and add to each allocation's systematic utility an independent Gumbel shock of scale $\sigma > 0$. The probability that A chooses a is then the binary logit (McFadden, 1974; Train, 2009)

$$f_\theta(Y = 1 \mid X) = \frac{\exp\{u(x_{g,a})/\sigma\}}{\exp\{u(x_{g,a})/\sigma\} + \exp\{u(x_{g,b})/\sigma\}} = \Lambda\left(\frac{u(x_{g,a}) - u(x_{g,b})}{\sigma}\right), \tag{12}$$

where $\Lambda(z) = 1/(1+e^{-z})$ and $\theta = (\lambda, \sigma)$ gathers the theory's preference parameters λ (the γ's) with the scale. The scale governs how noisy choice is: as σ falls the higher-utility allocation is chosen with greater probability, so $1/\sigma$ is a choice sensitivity. Because the theories of Section 1.3 nest, so do their forecast families, and the no-situation baseline of Section 1.1 lies below all of them.

A heterogeneity structure determines whether and how θ varies across subjects, and supplies the rungs of the individual-level ladder (Section 2.4). The *representative agent* gives every subject the common parameter θ. A K*-type* model assigns each subject to one of K latent types, with type parameters $\theta^1, \dots, \theta^K$ and population shares $\pi^1, \dots, \pi^K$ ($\pi^k \geq 0$, $\sum_k \pi^k = 1$), and forecasts by the mixture

$$f_K(Y = 1 \mid X) = \sum_{k=1}^{K} \pi^k\, f_{\theta^k}(Y = 1 \mid X), \tag{13}$$

which returns the representative agent at $K = 1$. The *mixed logit* is the continuum limit, drawing each subject's parameter from a distribution G rather than from a finite support (Revelt and Train, 1998). Section 2 sets out how each structure is estimated and how a held-out subject's forecast is formed from her own past choices.

2 Estimation strategy

The empirical exercise estimates the population objects of Section 1 and reports them with the one-sided semantics that Remark 1.1 licenses. We describe the design in the order the objects enter the completeness ratios — the cross-validation protocol, the two baselines and

[6]The inequality indicators are redundant given the payoffs; we retain them to keep the link between the situation and the social-preference models transparent.

the situation oracle, the frontier ceiling, and the heterogeneity ladder — and close with the inference protocol, which was fixed before the replication-arm and quadratic-loss reruns were computed (the repository's version history documents the ordering).

2.1 Cross-validation design and estimand

All reported risks are held-out log losses under subject-stratified five-fold cross-validation: each subject's choices are partitioned across the five folds, so every subject appears in every fold. The generalisation target is a new choice by an already-observed subject on the fixed situation grid — the information structure under which the K-type and mixed-logit forecasts are defined, since a held-out subject's posterior is formed from her own training-fold choices only (cf. the conditional–unconditional distinction in Krueger et al., 2021). Holding out whole subjects would answer a different (transfer) question, which we do not study. Every random source is seeded; all results are exactly reproducible.

2.2 Baselines and the situation oracle

The representative-agent scale runs from the unconditional training-fold choice frequency (baseline $H(Y)$) to the situation oracle $H(Y \mid X)$, estimated by per-situation choice frequencies shrunk toward the grand mean, with the pseudo-count chosen by internal cross-validation; the chosen penalty is reported as a data-sufficiency diagnostic. With 174 choices per situation the cells are well populated and the chosen shrinkage is small. The individual-level scale shares that same unconditional baseline $H(Y)$ — the two levels differ only in the oracle — and replaces the situation oracle with the frontier ceiling described next.

2.3 The ML frontier

The individual-level ceiling is the best held-out risk over a set of pooling predictors fit on the subject-by-situation panel: second-order factorisation machines (Rendle, 2010) on subject and situation one-hots plus the structural design, logistic low-rank matrix completion (cf. Candès and Recht, 2009), a subject-embedding model (per-subject ridge-penalised preference vectors in the full design space), and gradient boosting with subject identity as an encoded categorical. The frontier is model-agnostic: its feature space is always the full structural design, whichever theory is being scored. Within each training set, candidates compete on a validation slice (one quarter of the training rows, subject-stratified); the winner is refit on the full training data and predicts the test fold, so the winner is never selected on rows whose risk is reported. By Remark 1.1, every individual-level completeness number reported against the frontier is an *upper bound* on true individual-level completeness, and orderings across rungs are invariant to where the ceiling sits. The factorisation-machine configurations bracket a region located in a preliminary sweep of the main panel; we disclose this pre-screening, and the fold-seed replication in the appendix shows the resulting ceiling is insensitive to it.

2.4 The heterogeneity ladder

Each theory is estimated at three heterogeneity structures. The representative agent f_1 is a conditional logit fit by maximum likelihood (McFadden, 1974). The K-type rungs f_K, $K = 2, \ldots, 10$, are latent-class logits fit by expectation–maximisation (Dempster et al., 1977; McLachlan and Peel, 2000) with four independent restarts, keeping the highest-likelihood run; a held-out subject's forecast is her posterior-weighted mixture of type forecasts, the posterior formed from training-fold choices only. The mixed logit f_G (Revelt and Train, 1998) has diagonal-Gaussian mixing estimated by maximum simulated likelihood (at least 1,000 Monte Carlo draws) from ten starting points with disjoint seed blocks, keeping the run with the highest training simulated likelihood. We select restarts by likelihood rather than by optimiser exit flags because the latter proved unreliable in this application: a run flagged as converged can sit at an inferior local optimum, and the appendix reports per-fold restart-likelihood tables for both estimators together with the sensitivity of the headline contrast to restart depth and to the number of simulation draws. Estimation error in f_G can only *understate* the gaps reported below, while under-optimised K-type rungs would overstate them — hence the asymmetric attention.

2.5 Inference

The paper makes one prespecified inferential claim: the comparison of the three-type rung with the mixed logit, on the paired per-subject held-out loss difference. Because the panel is a complete grid, the subject-block statistic is exactly a mean of 174 independent per-subject mean-loss differences, so the headline interval is the paired t (173 degrees of freedom); a subject-block percentile bootstrap (the same resampling plan for every rung, preserving pairing) is reported as confirmation, and all completeness intervals come from the same joint resampling. All intervals are conditional on the fitted per-fold forecast rules: they quantify sampling variation over new subjects from the same population on the fixed grid, not re-estimation variability of the rules themselves; the fold-seed replication bounds the latter empirically, and unconditional cross-validation inference in the sense used by Fudenberg et al. (2026) for i.i.d. observations is not available for clustered panels with subject-conditioned forecasts. Gaps at other K are reported as point estimates with interval upper bounds on the share-of-gain scale; we make no equivalence claims, and the derived summary K^* — the smallest K whose gap interval covers zero — is read as a power-relative quantity, accompanied by the design's minimum detectable effect. The quadratic-loss (Brier) rerun (d'Eon et al., 2024) and the in-sample selection criteria of the mixture literature are reported in the appendix.

References

Gary E. Bolton and Axel Ockenfels. ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, 90(1):166–193, 2000.

Adrian Bruhin, Ernst Fehr, and Daniel Schunk. The many faces of human sociality: Uncovering the distribution and stability of social preferences. *Journal of the European Economic Association*, 17(4):1025–1069, 2019.

Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

Gary Charness and Matthew Rabin. Understanding social preferences with simple tests. *Quarterly Journal of Economics*, 117(3):817–869, 2002.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1):1–22, 1977.

Greg d'Eon, Sophie Greenwood, Kevin Leyton-Brown, and James R. Wright. How to evaluate behavioral models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)*, pages 9636–9644, 2024.

Ernst Fehr and Klaus M. Schmidt. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114(3):817–868, 1999.

Drew Fudenberg, Jon Kleinberg, Annie Liang, and Sendhil Mullainathan. Measuring the completeness of economic models. *Journal of Political Economy*, 130(4):956–990, 2022.

Drew Fudenberg, Wayne Gao, and Annie Liang. How flexible is that functional form? Quantifying the restrictiveness of theories. *Review of Economics and Statistics*, 108(1): 194–209, 2026.

Rico Krueger, Michel Bierlaire, Ricardo A. Daziano, Taha H. Rashidi, and Prateek Bansal. Evaluating the predictive abilities of mixed logit models with unobserved inter- and intra-individual heterogeneity. *Journal of Choice Modelling*, 41:100323, 2021.

David K. Levine. Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics*, 1(3):593–622, 1998.

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In Paul Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, 1974.

Geoffrey J. McLachlan and David Peel. *Finite Mixture Models*. Wiley, 2000.

Steffen Rendle. Factorization machines. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, pages 995–1000, 2010.

David Revelt and Kenneth Train. Mixed logit with repeated choices: Households' choices of appliance efficiency level. *Review of Economics and Statistics*, 80(4):647–657, 1998.

Kenneth E. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2 edition, 2009.