RES Easter School: Behavioural Economics
Brasenose College Oxford, 22-25 March 2015

Structural and nonparametric econometrics

Vincent P. Crawford

**University of Oxford, All Souls College,
and University of California, San Diego**

(thanks to Miguel Costa-Gomes, Bruno Broseta, Laura Blow, Ian Crawford, and Juanjuan Meng)

Revised 19 March 2015

Structural econometrics of decisions and search in games

The first part of this lecture concerns structural econometric analyses of strategic thinking in games, using maximum likelihood error-rate models of subjects' decisions and their searches for hidden payoff information.

The material for this first part is drawn mostly from:

- Costa-Gomes, Crawford, and Broseta, "Cognition and Behavior in Normal-Form Games: An Experimental Study," (2001 *Econometrica*; "CGCB") and
- Costa-Gomes and Crawford, "Cognition and Behavior in Two-Person Guessing Games: An Experimental Study," (2006 *American Economic Review*; "CGC"),

which build on:

- Harless and Camerer, "The Predictive Utility of Generalized Expected Utility Theories (1994 *Econometrica*) and
- El-Gamal and Grether, "Are People Bayesian? Uncovering Behavioral Strategies," (1995 *J. American Statistical Association*).

Recall that in most experiments that study strategic thinking, game-theoretically naïve subjects play series of different but related games with randomly, anonymously paired partners and no feedback.

The goal is to suppress learning and repeated-game effects, to elicit subjects' initial responses to each game, uncontaminated by learning.

Recall that in most experiments that study strategic thinking, game-theoretically naïve subjects play series of different but related games with randomly, anonymously paired partners and no feedback.

The goal is to suppress learning and repeated-game effects, to elicit subjects' initial responses to each game, uncontaminated by learning.

CGCB's subjects played a series of 18 matrix games, chosen to separate nonstrategic from strategic types as much as possible.

(CGCB's results detected strategic thinking without discriminating among strategic types as clearly as desired, while ruling out some other types.)

CGC's subjects played a series of 16 two-person guessing games, chosen to further separate strategic types as much as possible.

Recall that in most experiments that study strategic thinking, game-theoretically naïve subjects play series of different but related games with randomly, anonymously paired partners and no feedback.

The goal is to suppress learning and repeated-game effects, to elicit subjects' initial responses to each game, uncontaminated by learning.

CGCB's subjects played a series of 18 matrix games, chosen to separate nonstrategic from strategic types as much as possible.

(CGCB's results detected strategic thinking without discriminating among strategic types as clearly as desired, while ruling out some other types.)

CGC's subjects played a series of 16 two-person guessing games, chosen to further separate strategic types as much as possible.

In each case the design monitored subjects' searches for hidden but freely accessible payoff information along with their decisions, with the goal of more precisely estimating their decision rules.

The econometric problem is characterizing subjects' heterogeneous decision and search sequences in a huge space of possible sequences.

CGCB and CGC did this by assuming (with testing) that each subject's decisions and searches are determined in all games, up to errors, by a single strategic decision rule or "type" (*not* a private-information variable).

The econometric problem is characterizing subjects' heterogeneous decision and search sequences in a huge space of possible sequences.

CGCB and CGC did this by assuming (with testing) that each subject's decisions and searches are determined in all games, up to errors, by a single strategic decision rule or "type" (*not* a private-information variable).

The types are a basis for the space of decision and search sequences, whose structure allows a tractable description of behavior and makes it meaningful to ask how subjects' decisions and searches are related.

The econometric problem is characterizing subjects' heterogeneous decision and search sequences in a huge space of possible sequences.

CGCB and CGC did this by assuming (with testing) that each subject's decisions and searches are determined in all games, up to errors, by a single strategic decision rule or "type" (*not* a private-information variable).

The types are a basis for the space of decision and search sequences, whose structure allows a tractable description of behavior and makes it meaningful to ask how subjects' decisions and searches are related.

CGCB estimated a mixture model of types, and CGC then estimated subjects' types subject by subject. (An early version of CGCB estimated subject by subject estimates, with almost the same results.)

Both approaches allow for rich heterogeneity, but in a mixture model a subject's behavior is treated as evidence even about others' types.

An intermediate approach is to estimate a clustering or latent class model, adding classes as long as they add enough to fit to justify it.

Why not estimate the distribution of rules without imposing a structure?

Unrestricted lists of types allow overfitting via types that “just happen” to do what subjects did in the sample.

A worthy alternative to equilibrium must be a general decision rule whose implications in new games are clear, not just a list of predicted decisions.

And because a type’s search implications depend not only on what decisions it implies but why, such lists give us no way to predict search.

CGCB's and CGC's types all build in risk-neutrality and (except for CGCB's *Altruistic*) rule out social preferences:

- *Altruistic* (CGCB only), which maximizes the sum of its own and other's payoffs over all possible decision combinations

CGCB's and CGC's types all build in risk-neutrality and (except for CGCB's *Altruistic*) rule out social preferences:

- *Altruistic* (CGCB only), which maximizes the sum of its own and other's payoffs over all possible decision combinations
- *Pessimistic* ("maximin"; CGCB only), which (without randomizing) maximizes its minimum payoff over other's possible decisions
- *Optimistic* ("maximax"; CGCB only), which maximizes its maximum payoff over other's decisions

CGCB's and CGC's types all build in risk-neutrality and (except for CGCB's *Altruistic*) rule out social preferences:

- *Altruistic* (CGCB only), which maximizes the sum of its own and other's payoffs over all possible decision combinations
- *Pessimistic* ("maximin"; CGCB only), which (without randomizing) maximizes its minimum payoff over other's possible decisions
- *Optimistic* ("maximax"; CGCB only), which maximizes its maximum payoff over other's decisions
- *Naïve* (CGC's *L1*), which best responds to "*L0*" beliefs that assign equal probabilities to the other player's feasible decisions (with flexible error structure and type distribution, *L0*'s frequency estimates are 0)
- *L2*, which best responds to *Naïve* (CGC's *L1*)
- *L3* (CGC only), which best responds to *L2*

CGCB's and CGC's types all build in risk-neutrality and (except for CGCB's *Altruistic*) rule out social preferences:

- *Altruistic* (CGCB only), which maximizes the sum of its own and other's payoffs over all possible decision combinations
- *Pessimistic* ("maximin"; CGCB only), which (without randomizing) maximizes its minimum payoff over other's possible decisions
- *Optimistic* ("maximax"; CGCB only), which maximizes its maximum payoff over other's decisions
- *Naïve* (CGC's *L1*), which best responds to "*L0*" beliefs that assign equal probabilities to the other player's feasible decisions (with flexible error structure and type distribution, *L0*'s frequency estimates are 0)
- *L2*, which best responds to *Naïve* (CGC's *L1*)
- *L3* (CGC only), which best responds to *L2*
- *D1* (*D2*), which does one round (two) of deleting decisions dominated by pure decisions and best responds to a uniform prior over the other player's remaining decisions (*D1*, like *L2*, is *k*-rationalizable)

CGCB's and CGC's types all build in risk-neutrality and (except for CGCB's *Altruistic*) rule out social preferences:

- *Altruistic* (CGCB only), which maximizes the sum of its own and other's payoffs over all possible decision combinations
- *Pessimistic* ("maximin"; CGCB only), which (without randomizing) maximizes its minimum payoff over other's possible decisions
- *Optimistic* ("maximax"; CGCB only), which maximizes its maximum payoff over other's decisions
- *Naïve* (CGC's *L1*), which best responds to "*L0*" beliefs that assign equal probabilities to the other player's feasible decisions (with flexible error structure and type distribution, *L0*'s frequency estimates are 0)
- *L2*, which best responds to *Naïve* (CGC's *L1*)
- *L3* (CGC only), which best responds to *L2*
- *D1* (*D2*), which does one round (two) of deleting decisions dominated by pure decisions and best responds to a uniform prior over the other player's remaining decisions (*D1*, like *L2*, is *k*-rationalizable)
- *Equilibrium*, which makes its (unique) equilibrium decision
- *Sophisticated*, which best responds to the probability distributions of others' decisions, proxied by the observed population frequencies

CGCB's maximum likelihood error-rate analysis

- In CGCB's mixture model, each subject's type is drawn from a common prior population distribution, whose types are specified a priori.

CGCB’s maximum likelihood error-rate analysis

- In CGCB’s mixture model, each subject’s type is drawn from a common prior population distribution, whose types are specified a priori.

- The model assumes that in each game, a subject’s type determines his information search, represented by categorical variables that reflect its compliance with each type’s Occurrence and Adjacency implications (as discussed elsewhere in slides), with errors, and his type and search then determine his decision, with uniform random errors (to avoid bias).

CGCB's maximum likelihood error-rate analysis

- In CGCB's mixture model, each subject's type is drawn from a common prior population distribution, whose types are specified a priori.

- The model assumes that in each game, a subject's type determines his information search, represented by categorical variables that reflect its compliance with each type's Occurrence and Adjacency implications (as discussed elsewhere in slides), with errors, and his type and search then determine his decision, with uniform random errors (to avoid bias).

- The goal is to infer the distribution of types (and implicitly each subject's type, which can be inferred by conditioning on the history) from subjects' decision and search sequences across all games.

- CGCB's maximum likelihood estimation does this inference by comparing subjects' decisions or decisions and searches over all games, with each type's predicted decisions and searches, taking type-*k* behavior as evidence for type *k* only to the extent that the estimated error rates suggest it was more likely than non-type *k* behavior.

- Conditional on type and game size, decision and search errors can be correlated in a given game, with decision error rates conditional on type and search compliance.

- Decision and search errors are assumed to be i.i.d. across games and subjects: A general joint probability distribution, except for constraints on how game size matters and how search compliance is defined.

CGCB's estimates from decisions alone

- *Naïve* and *Optimistic* always make the same decisions; they are lumped together pending the search analysis, where separated.
- *L2* and *Sophisticated* decisions are separated, weakly, in only one game for Columns; but identified in the pooled Row and Column data.
- Any two other types make different decisions in at least 2/18 games for each player role; strategic and nonstrategic types strongly separated.

CGCB’s estimates from decisions alone

- *Naïve* and *Optimistic* always make the same decisions; they are lumped together pending the search analysis, where separated.
- *L2* and *Sophisticated* decisions are separated, weakly, in only one game for Columns; but identified in the pooled Row and Column data.
- Any two other types make different decisions in at least 2/18 games for each player role; strategic and nonstrategic types strongly separated.
- Maximum likelihood yields consistent estimates of the model’s 7 independent type probabilities and 8 type-dependent error rates.

CGCB's estimates from decisions alone

- *Naïve* and *Optimistic* always make the same decisions; they are lumped together pending the search analysis, where separated.
- *L2* and *Sophisticated* decisions are separated, weakly, in only one game for Columns; but identified in the pooled Row and Column data.
- Any two other types make different decisions in at least 2/18 games for each player role; strategic and nonstrategic types strongly separated.
- Maximum likelihood yields consistent estimates of the model's 7 independent type probabilities and 8 type-dependent error rates.
- Few subjects are estimated to be *Equilibrium* and none *Sophisticated*.

CGCB's estimates from decisions alone

- *Naïve* and *Optimistic* always make the same decisions; they are lumped together pending the search analysis, where separated.
- *L2* and *Sophisticated* decisions are separated, weakly, in only one game for Columns; but identified in the pooled Row and Column data.
- Any two other types make different decisions in at least 2/18 games for each player role; strategic and nonstrategic types strongly separated.
- Maximum likelihood yields consistent estimates of the model's 7 independent type probabilities and 8 type-dependent error rates.
- Few subjects are estimated to be *Equilibrium* and none *Sophisticated*.
- Many subjects are estimated to be *Naïve* (*L1*) or *Optimistic*.

CGCB's estimates from decisions alone

- *Naïve* and *Optimistic* always make the same decisions; they are lumped together pending the search analysis, where separated.
- *L2* and *Sophisticated* decisions are separated, weakly, in only one game for Columns; but identified in the pooled Row and Column data.
- Any two other types make different decisions in at least 2/18 games for each player role; strategic and nonstrategic types strongly separated.
- Maximum likelihood yields consistent estimates of the model's 7 independent type probabilities and 8 type-dependent error rates.
- Few subjects are estimated to be *Equilibrium* and none *Sophisticated*.
- Many subjects are estimated to be *Naïve* (*L1*) or *Optimistic*.
- Many others are estimated to be the 2-rationalizable types *L2* or *D1*.

CGCB's estimates from decisions alone

- *Naïve* and *Optimistic* always make the same decisions; they are lumped together pending the search analysis, where separated.
- *L2* and *Sophisticated* decisions are separated, weakly, in only one game for Columns; but identified in the pooled Row and Column data.
- Any two other types make different decisions in at least 2/18 games for each player role; strategic and nonstrategic types strongly separated
- Maximum likelihood yields consistent estimates of the model's 7 independent type probabilities and 8 type-dependent error rates
- Few subjects are estimated to be *Equilibrium* and none *Sophisticated*.
- Many subjects are estimated to be *Naïve* (*L1*) or *Optimistic*.
- Many others are estimated to be the 2-rationalizable types *L2* or *D1*.

  (Even though *L2* is weakly separated from *Sophisticated*, its better fit in one game is amplified to a large lead by lower estimated error rates.)

  (These types do well because they reproduce many subjects' tendency to play equilibrium in simple games but switch to *Naïve* (*L1*) in others.)

CGCB's estimates from decisions and search

- There is strong separation of search implications across three groups of types: (i) *Altruistic*; (ii) *Pessimistic*, *Naïve* (*L1*), or *Optimistic*; and (iii) *L2*, *D1*, *D2*, *Equilibrium*, or *Sophisticated*.

- There is also some separation within groups, e.g. *L2* from *D1*.

CGCB's estimates from decisions and search

- There is strong separation of search implications across three groups of types: (i) *Altruistic*; (ii) *Pessimistic*, *Naïve* (*L1*), or *Optimistic*; and (iii) *L2*, *D1*, *D2*, *Equilibrium*, or *Sophisticated*.

- There is also some separation within groups, e.g. *L2* from *D1*.

- Maximum likelihood again yields consistent parameter estimates.

- Again, type-*k* behavior is taken as evidence for type *k* to the extent that estimated error rates suggest it was more likely than other behavior; but now the search terms in the likelihood, which are convex in search compliance, favor types for which compliance is concentrated on particular levels, lowering estimated error rates.

  (Compliance of course should be concentrated on *high* levels; estimates done unconstrained as a check confirm that.)

- Estimates for the model of decisions and information search generally confirm the type estimates from decisions alone, with some changes. Incorporating types' cognitive implications into an error rate analysis yields a coherent account of subjects' behavior and better predictions.

- Estimates for the model of decisions and information search generally confirm the type estimates from decisions alone, with some changes. Incorporating types' cognitive implications into an error rate analysis yields a coherent account of subjects' behavior and better predictions.

- *Naïve* (*L1*) and *L2* now have the largest frequencies, each around 45%; and *D1* has disappeared.

  The shift toward *Naïve* (*L1*), mainly at the expense of *Optimistic* and *D1*, happens because *Naïve* (*L1*) search compliance explains more of the variation in subjects' behavior than *Optimistic* search compliance, which is too unrestrictive to be useful, or *D1* compliance, which is more restrictive than *Naïve* (*L1*)'s but less correlated with subjects' behavior.

- Estimates for the model of decisions and information search generally confirm the type estimates from decisions alone, with some changes. Incorporating types' cognitive implications into an error rate analysis yields a coherent account of subjects' behavior and better predictions.

- *Naïve* (*L1*) and *L2* now have the largest frequencies, each around 45%; and *D1* has disappeared.

  The shift toward *Naïve* (*L1*), mainly at the expense of *Optimistic* and *D1*, happens because *Naïve* (*L1*) search compliance explains more of the variation in subjects' behavior than *Optimistic* search compliance, which is too unrestrictive to be useful, or *D1* compliance, which is more restrictive than *Naïve* (*L1*)'s but less correlated with subjects' behavior.

- *Naive* (*L1*) and *L2* have high search compliance, error rates that decrease with higher compliance and are low with high compliance.

- Estimates for the model of decisions and information search generally confirm the type estimates from decisions alone, with some changes. Incorporating types' cognitive implications into an error rate analysis yields a coherent account of subjects' behavior and better predictions.

- *Naïve* (*L1*) and *L2* now have the largest frequencies, each around 45%; and *D1* has disappeared.

  The shift toward *Naïve* (*L1*), mainly at the expense of *Optimistic* and *D1*, happens because *Naïve* (*L1*) search compliance explains more of the variation in subjects' behavior than *Optimistic* search compliance, which is too unrestrictive to be useful, or *D1* compliance, which is more restrictive than *Naïve* (*L1*)'s but less correlated with subjects' behavior.

- *Naive* (*L1*) and *L2* have high search compliance, error rates that decrease with higher compliance and are low with high compliance.

- *D1* has fairly high compliance and high error rates that usually decrease with compliance.

- *Altruistic* and *Equilibrium* have low compliance and error rates that usually decrease with compliance.

CGC's estimates from decisions alone

● As noted previously, CGC's design, with 200 to 800 possible exact (within 0.5) guesses in 16 different games, yields very strong separation of types' guesses.

(a_i,b_i, and p_i are a player's lower limit, upper limit, and target, and a_j,b_j, and p_j are his partner's.)

Types' guesses in the 16 games, in (randomized) order played

Game	a_i	b_i	p_i	a_j	b_j	p_j	***L1***	***L2***	***L3***	***D1***	***D2***	***Eq***	***So***
1	100	900	1.5	300	500	0.7	600	525	630	600	611.25	750	630
2	300	900	1.3	300	500	1.5	520	650	650	617.5	650	650	650
3	300	900	1.3	300	900	1.3	780	900	900	838.5	900	900	900
4	300	900	0.7	100	900	1.3	350	546	318.5	451.5	423.15	300	420
5	100	500	1.5	100	500	0.7	450	315	472.5	337.5	341.25	500	375
6	100	500	0.7	100	900	0.5	350	105	122.5	122.5	122.5	100	122
7	100	500	0.7	100	500	1.5	210	315	220.5	227.5	227.5	350	262
8	300	500	0.7	100	900	1.5	350	420	367.5	420	420	500	420
9	300	500	1.5	300	900	1.3	500	500	500	500	500	500	500
10	300	500	0.7	100	900	0.5	350	300	300	300	300	300	300
11	100	500	1.5	100	900	0.5	500	225	375	262.5	262.5	150	300
12	300	900	1.3	300	900	1.3	780	900	900	838.5	900	900	900
13	100	900	1.3	300	900	0.7	780	455	709.8	604.5	604.5	390	695
14	100	900	0.5	300	500	0.7	200	175	150	200	150	150	162
15	100	900	0.5	100	500	0.7	150	175	100	150	100	100	132
16	100	900	0.5	100	500	1.5	150	250	112.5	162.5	131.25	100	187

- Of CGC's 88 subjects, 43's guesses complied *exactly* with one type's guesses in from 7 to 16 games: 20 *L1*, 12 *L2*, 3 *L3*, and 8 *Equilibrium*.

- Those subjects' types can confidently be identified from their guessing "fingerprints", without econometrics (with qualification for *Equilibrium*).

  (By contrast, there are usually many possible reasons for choosing one of a few strategies in a small matrix game; and even in Nagel's games, rules as cognitively disparate as *Dk* and *Lk*+1 make identical guesses.)

- Further, because CGC's definitions of level-*k* types builds in risk-neutral, self-interested rationality, we know that the deviations from equilibrium of subjects with high exact compliance with level-*k* types are caused not by irrationality, risk aversion, altruism, spite, or confusion, but by their simplified model of others.

CGC's maximum likelihood error-rate analysis

- CGC's other 45 subjects' guesses conformed less exactly to a type; estimating their types requires econometrics.

- CGC estimated a model generally similar to CGCB's, again with types specified a priori, but subject by subject.

- The model again assumes that in each game, a subject's type determines his information search, with errors, and his type and search then determine his decision, now with logit rather than uniform errors.

- The goal is again to infer each subject's type from his decisions and searches across all games.

- CGC's maximum likelihood error-rate analysis does this inference by comparing subjects' decisions, or decisions and searches, over all games, with each type's predicted decisions and searches, taking type-*k* behavior as evidence for type *k* only to the extent that the estimated error rates suggest it was more likely than non-type *k* behavior.

- CGC's maximum likelihood error-rate analysis does this inference by comparing subjects' decisions, or decisions and searches, over all games, with each type's predicted decisions and searches, taking type-*k* behavior as evidence for type *k* only to the extent that the estimated error rates suggest it was more likely than non-type *k* behavior.

- Conditional on type, decision and search errors can be correlated in a given game, with decision error rates conditional on type and search compliance.

- Decision and search errors are assumed to be i.i.d. across games.

- CGC's maximum likelihood error-rate analysis does this inference by comparing subjects' decisions, or decisions and searches, over all games, with each type's predicted decisions and searches, taking type-*k* behavior as evidence for type *k* only to the extent that the estimated error rates suggest it was more likely than non-type *k* behavior.

- Conditional on type, decision and search errors can be correlated in a given game, with decision error rates conditional on type and search compliance.

- Decision and search errors are assumed to be i.i.d. across games.

- Because of the very high sample frequency of exact guesses, CGC allowed "spike-logit" errors: In each game, a subject makes his type's guess exactly (within 0.5) with probability 1- ε and otherwise makes logit errors (extra likelihood credit for exact guesses, whose weight is discontinuously higher than guesses that are close but not within 0.5).

Subject *i*'s log-likelihood for guesses alone reduces to:

(7) $$(G-n^{ik})\ln(G-n^{ik})+n^{ik}\ln(n^{ik})+\sum_{g\in N^{ik}}\ln d_g^k(R_g^i(x_g^i),\lambda)-G\ln G,$$

where *g* indexes games and *k* indexes types.

The first two terms concern exact guesses; $d_g^k(R_g^i(x_g^i),\lambda)$ is the standard logit term for non-exact guesses, with deviation costs measured using each type's beliefs; and λ is the logit precision.

Subject *i*'s log-likelihood for guesses alone reduces to:

(7) $$(G-n^{ik})\ln(G-n^{ik})+n^{ik}\ln(n^{ik})+\sum_{g\in N^{ik}}\ln d_g^k(R_g^i(x_g^i),\lambda)-G\ln G,$$

where *g* indexes games and *k* indexes types.

The first two terms concern exact guesses; $d_g^k(R_g^i(x_g^i),\lambda)$ is the standard logit term for non-exact guesses, with deviation costs measured using each type's beliefs; and *λ* is the logit precision.

The maximum likelihood estimate of ε is n^{ik} /*G*, the sample frequency of subject *i*'s non-exact guesses for type *k*.

The maximum likelihood estimate of *λ* is the standard logit precision, restricted to non-exact guesses.

The maximum likelihood estimate of the subject's type *k* maximizes (7) over *k*, given the estimated ε and *λ*, trading off the count of exact guesses against the logit cost of deviations.

CGC's estimates from guesses alone

- The hypothesis that $\varepsilon = 1$ is rejected for all but 7 of 88 subjects: the spike is necessary.

- The hypothesis that $\lambda = 0$ (payoff-insensitivity) is rejected for 34 subjects: logit errors significantly improve the fit over a spike-uniform model like CGCB's for only 39% of the subjects, suggesting that most "errors" are either cognitive or due to misspecification.

- The hypothesis that $\{\lambda = 0 \text{ and } \varepsilon = 1\}$ is rejected at the 5% level for all but 10 subjects: the model does better than random for 89% of them.

- Estimation from guesses alone yields type estimates as in column 3 of Table 1: 43 *L1*, 20 *L2*, 3 *L3*, 5 *D1*, 14 *Equilibrium*, and 3 *Sophisticated*.

  (Some of these estimates are called into question by CGC's specification test, discussed below; see Table 1's columns 4 and 5).

- Unlike the often-suggested interpretation of previous guessing results that subjects are performing finitely iterated dominance, separating *Lk* from *Dk-1* reveals that *Dk* types don't exist in any significant numbers.

  (Results for R/TS subjects not discussed here suggest that people find iterated dominance highly unnatural—as opposed to *Lk*'s iterated best resonses—and so respect finitely iterated dominance without explicitly performing it.)

- *Sophisticated*, which is clearly separated from *Equilbrium* here because few subjects play equilibrium strategies, also doesn't exist.

Table 1—Summary of Baseline and OB Subjects' Estimated Type Distributions

Type	Apparent from guesses	Econometric from guesses	Econometric from guesses, excluding random	Econometric from guesses, with specification test	Econometric from guesses and search, with specification test
L1	20	43	37	27	29
L2	12	20	20	17	14
L3	3	3	3	1	1
D1	0	5	3	1	0
D2	0	0	0	0	0
Eq.	8	14	13	11	10
Soph.	0	3	2	1	1
Unclassified	45	0	10	30	33

Note: The far-right-hand column includes 17 OB subjects classified by their econometric-from-guesses type estimates.

CGC's estimates from guesses and search

- The main problem in econometrically analyzing search is extracting signals from highly idiosyncratic, noisy searches, without a theoretical model that implies strong restrictions on how cognition drives search.

- CGC summarize a subject's compliance with a type's search implications in a game roughly by the density of the type's characteristic look-up sequence (as discussed previously) in the subject's look-up sequence.

- CGC further assume (simplifying CGCB) that, given type, errors in search and guesses are independent of each other and across games.

  (This simplifying assumption makes the log-likelihood separable across guesses and search, avoiding some complications.)

- To avoid stronger distributional assumptions with no theory to guide them, CGC discretized search compliance into three categories:

$$C_H \equiv [0.67,1.00],\ C_M \equiv [0.33,0.67],\text{ and } C_L \equiv [0,0.33].$$

Subject *i*'s guesses-and-search log-likelihood is:

$$\sum_c \left[m_c^{isk} \ln(\zeta_c) + (m_c^{isk} - n_c^{isk}) \ln(1-\varepsilon) + n^{isk} \ln(\varepsilon) + \sum_{g \in N_c^{isk}} \ln d_g^k (R_g^i(x_g^i), \lambda) \right] \equiv$$

$$(G-n^{ik})\ln(G-n^{ik}) + n^{ik}\ln(n^{ik}) + \sum_{g \in N^{ik}} \ln d_g^k(R_g^i(x_g^i),\lambda) - G\ln G + \sum_c \left[m_c^{isk} \ln m_c^{isk} \right] - 2G\ln G,$$

m_c^{isk} is the number of games for which subject *i* has type-*k* style-*s* compliance *c* ("style" is where the subject's relevant look-ups are).

The search term is convex in the m_c^{isk} and therefore favors types for which compliance varies less across games, because such types "explain" search behavior better, as in CGCB.

The maximum-likelihood estimates of ε and ζ_c, given *k* and *s*, are n^{ik} /*G* and m_c^{isk} / G, the sample frequencies with which subject *i*'s adjusted guesses are non-exact for that *k* and *i* has compliance *c* for that *k* and *s*.

The maximum likelihood estimate of *λ* is the standard logit precision.

- The maximum likelihood estimate of subject *i*'s type *k* maximizes the above log-likelihood over *k* and *s*, given the estimated ε and *λ*.

- Most guesses-and-search type estimates, especially for subjects whose guess fingerprints were clear, reaffirm guesses-only estimates, including the absence of significant numbers of subjects of types other than *L1*, *L2*, *Equilibrium*, or hybrids of *L3* or *Equilibrium*.

- Incorporating search does refine and sharpen conclusions in some ways, and a few subjects' type estimates change.

- And most subjects' types can be more precisely identified by decisions and search than by decisions or search alone (Table 7B):

- The search part of the likelihood has weight only about 1/6 of the precise predictions than our theory of guesses—a necessary evil, given the noisiness and idiosyncrasy of search behavior.

TABLE 1—SUMMARY OF BASELINE AND OB SUBJECTS' ESTIMATED TYPE DISTRIBUTIONS

Type	Apparent from guesses	Econometric from guesses	Econometric from guesses, excluding random	Econometric from guesses, with specification test	Econometric from guesses and search, with specification test
L1	20	43	37	27	29
L2	12	20	20	17	14
L3	3	3	3	1	1
D1	0	5	3	1	0
D2	0	0	0	0	0
Eq.	8	14	13	11	10
Soph.	0	3	2	1	1
Unclassified	45	0	10	30	33

Note: The far-right-hand column includes 17 OB subjects classified by their econometric-from-guesses type estimates.

CGC's specification test

For the 45 subjects whose guesses conformed less exactly to one of CGC's types, there is room for doubt about whether CGC's specification omits relevant types and/or overfits by including irrelevant types.

CGC conducted a specification test, comparing the guesses-only likelihood of each subject's type estimate with those of estimates based on 88 *pseudotypes*, each constructed one subject's guesses.

If CGC had omitted a relevant type, say *L2*, the pseudotypes of subjects CGC now estimated to be *L2* would outperform their hypothetical non-*L2* estimated types, and would all make approximately the same guesses.

CGC found five such small *cluster*s involving a total of 11 subjects, diagnosed omitted types, and left those subjects unclassified in Table 1.

With regard to overfitting via irrelevant types, a credible type estimate for a subject should have higher likelihood than as many pseudotypes as it would at random: With 8 types, assuming approximately i.i.d. likelihoods, it should have higher likelihood than 87/8 ≈ 11 pseudotypes.

Some subjects' type estimates do not pass this test, and so are left unclassified in columns 5 and 6 of CGC's Table 1.

With this classification econometric estimates of subjects' types are concentrated on *L1*, *L2*, *L3*, and *Equilibrium* in roughly the proportions as for subjects whose types are apparent from their guesses alone.

TABLE 1—SUMMARY OF BASELINE AND OB SUBJECTS' ESTIMATED TYPE DISTRIBUTIONS

Type	Apparent from guesses	Econometric from guesses	Econometric from guesses, excluding random	Econometric from guesses, with specification test	Econometric from guesses and search, with specification test
L1	20	43	37	27	29
L2	12	20	20	17	14
L3	3	3	3	1	1
D1	0	5	3	1	0
D2	0	0	0	0	0
Eq.	8	14	13	11	10
Soph.	0	3	2	1	1
Unclassified	45	0	10	30	33

Note: The far-right-hand column includes 17 OB subjects classified by their econometric-from-guesses type estimates.

Revealed reference-dependent preference and cabdrivers' labor supply

Recall that Crawford and Meng (2011 *AER*) adapted Kőszegi and Rabin's (2006 *QJE*) model of reference-dependent preferences to continue the Camerer et al. (1997 *QJE*)-Farber (2005 *JPE*, 2008 *AER*) debate on cabdrivers' labor supply. As applied to cabdrivers' labor supply

- A driver's preferences reflect both the standard consumption utility of income and leisure and reference-dependent "gain-loss" utility, with their relative importance tuned by an estimated parameter.

Revealed reference-dependent preference and cabdrivers' labor supply

Recall that Crawford and Meng (2011 *AER*) adapted Kőszegi and Rabin's (2006 *QJE*) model of reference-dependent preferences to continue the Camerer et al. (1997 *QJE*)-Farber (2005 *JPE*, 2008 *AER*) debate on cabdrivers' labor supply. As applied to cabdrivers' labor supply

- A driver's preferences reflect both the standard consumption utility of income and leisure and reference-dependent "gain-loss" utility, with their relative importance tuned by an estimated parameter.

  (The "only" deviation from a neoclassical model is adding changes in income and leisure to their levels in the domain of preferences.)

Revealed reference-dependent preference and cabdrivers' labor supply

Recall that Crawford and Meng (2011 *AER*) adapted Kőszegi and Rabin's (2006 *QJE*) model of reference-dependent preferences to continue the Camerer et al. (1997 *QJE*)-Farber (2005 *JPE*, 2008 *AER*) debate on cabdrivers' labor supply. As applied to cabdrivers' labor supply

- A driver's preferences reflect both the standard consumption utility of income and leisure and reference-dependent "gain-loss" utility, with their relative importance tuned by an estimated parameter.

  (The "only" deviation from a neoclassical model is adding changes in income and leisure to their levels in the domain of preferences.)

- A driver has a daily target for hours as well as income, and he is loss-averse in both dimensions, with working longer than the hours target a loss, just as earning less than the income target is.

Revealed reference-dependent preference and cabdrivers' labor supply

Recall that Crawford and Meng (2011 *AER*) adapted Kőszegi and Rabin's (2006 *QJE*) model of reference-dependent preferences to continue the Camerer et al. (1997 *QJE*)-Farber (2005 *JPE*, 2008 *AER*) debate on cabdrivers' labor supply. As applied to cabdrivers' labor supply

- A driver's preferences reflect both the standard consumption utility of income and leisure and reference-dependent "gain-loss" utility, with their relative importance tuned by an estimated parameter.

  (The "only" deviation from a neoclassical model is adding changes in income and leisure to their levels in the domain of preferences.)

- A driver has a daily target for hours as well as income, and he is loss-averse in both dimensions, with working longer than the hours target a loss, just as earning less than the income target is.
- Most importantly, the targets are endogenized by setting them equal to a driver's theoretical rational expectations of hours and income, in Kőszegi and Rabin's notion of "preferred personal equilibrium", operationalized via natural sample proxies.

Details

Treating each day separately as in previous such analyses, consider the preferences of a given driver during his shift on a given day.

I and H denote income earned and hours worked that day; I^r and H^r denote income and hours targets for the day.

Details

Treating each day separately as in previous such analyses, consider the preferences of a given driver during his shift on a given day.

I and H denote income earned and hours worked that day; I^r and H^r denote income and hours targets for the day.

Total utility, $V(I, H|I^r, H^r)$, is a weighted average of consumption utility $U_1(I) + U_2(H)$ and gain-loss utility $R(I, H|I^r, H^r)$, with weights $1 – \eta$ and η ($0 \leq \eta \leq 1$):

(1) $$V(I,H \mid I^r,H^r) = (1-\eta)(U_1(I)+U_2(H)) + \eta R(I,H \mid I^r,H^r)$$

Details

Treating each day separately as in previous such analyses, consider the preferences of a given driver during his shift on a given day.

I and *H* denote income earned and hours worked that day; I^r and H^r denote income and hours targets for the day.

Total utility, $V(I, H|I^r, H^r)$, is a weighted average of consumption utility $U_1(I) + U_2(H)$ and gain-loss utility $R(I, H|I^r, H^r)$, with weights $1 - \eta$ and η ($0 \leq \eta \leq 1$):

(1) $$V(I, H \mid I^r, H^r) = (1-\eta)(U_1(I) + U_2(H)) + \eta R(I, H \mid I^r, H^r)$$

where gain-loss utility

(2) $$R(I, H \mid I^r, H^r) = 1_{(I-I^r \leq 0)} \lambda(U_1(I) - U_1(I^r)) + 1_{(I-I^r > 0)}(U_1(I) - U_1(I^r)) + 1_{(H-H^r \geq 0)} \lambda(U_2(H) - U_2(H^r)) + 1_{(H-H^r < 0)}(U_2(H) - U_2(H^r)).$$

Kőszegi and Rabin's, Farber's, and Crawford and Meng's analyses make assumptions on functional structure that go beyond theory or evidence:

- Consumption utility is additively separable across income and hours, with $U_1(\cdot)$ increasing in I, $U_2(\cdot)$ decreasing in H, and both concave.

- Consumption utility has a particular, standard functional form.

- Gain-loss utility is additively separable, determined good by good by differences between goods' realized and target consumption utilities.

- Gain-loss utility is linear in good by good utility differences (constant sensitivity).

  (With constant sensitivity the previous assumption reduces to additive separability across goods of our u(q,q−r) utility function below.)

- Losses have a constant weight relative to gains, (the "coefficient of loss aversion", empirically ≈ 2 to 3), the same for income and hours.

Blow, Crawford, and Crawford (2015; "BCC") study nonparametric versions of a generalized Kőszegi-Rabin model, to learn to what extent the empirical success of applications is due to Kőszegi and Rabin's ancillary structural assumptions or reference-dependence per se.

Blow, Crawford, and Crawford (2015; “BCC”) study nonparametric versions of a generalized Kőszegi-Rabin model, to learn to what extent the empirical success of applications is due to Kőszegi and Rabin’s ancillary structural assumptions or reference-dependence per se.

BCC derive nonparametric necessary and sufficient conditions, in the revealed-preference tradition of Samuelson, Houthakker, and Afriat, for the existence of reference-dependent preferences that rationalize choice as in Kőszegi and Rabin’s model, without restricting functional structure.

Blow, Crawford, and Crawford (2015; "BCC") study nonparametric versions of a generalized Kőszegi-Rabin model, to learn to what extent the empirical success of applications is due to Kőszegi and Rabin's ancillary structural assumptions or reference-dependence per se.

BCC derive nonparametric necessary and sufficient conditions, in the revealed-preference tradition of Samuelson, Houthakker, and Afriat, for the existence of reference-dependent preferences that rationalize choice as in Kőszegi and Rabin's model, without restricting functional structure.

Given the revealed-preference tradition's reliance on rationality, such an analysis is possible, despite the non-neoclassical inclusion of changes as well as levels in the domain of preferences, only because the model is consistent with rationality in the broad sense of choice consistency.

The observable implications of reference-dependent preferences turn on:

- Whether "sensitivity" is constant ("sign-dependence") or diminishing (really, variable), and

- Whether reference points are unobservable or observable (really, modelable as known functions of the data, as via sample proxies).

The observable implications of reference-dependent preferences turn on:

- Whether "sensitivity" is constant ("sign-dependence") or diminishing (really, variable), and

- Whether reference points are unobservable or observable (really, modelable as known functions of the data, as via sample proxies).

BCC consider all four cases, but focus on the case of constant sensitivity and observable reference points, which seems most useful.

The observable implications of reference-dependent preferences turn on:

- Whether "sensitivity" is constant ("sign-dependence") or diminishing (really, variable), and

- Whether reference points are unobservable or observable (really, modelable as known functions of the data, as via sample proxies).

BCC consider all four cases, but focus on the case of constant sensitivity and observable reference points, which seems most useful.

They illustrate the results by nonparametrically analyzing the dataset on New York City cabdrivers studied by Farber and Crawford and Meng:

The observable implications of reference-dependent preferences turn on:

- Whether "sensitivity" is constant ("sign-dependence") or diminishing (really, variable), and

- Whether reference points are unobservable or observable (really, modelable as known functions of the data, as via sample proxies).

BCC consider all four cases, but focus on the case of constant sensitivity and observable reference points, which seems most useful.

They illustrate the results by nonparametrically analyzing the dataset on New York City cabdrivers studied by Farber and Crawford and Meng:

- The analysis rejects most of the ancillary separability and functional structure assumptions maintained in previous work.

- But reference-dependence still allows a plausible, rationality-based explanation of drivers' choices.

BCC focus on the leading case of multiple price-quantity observations for one consumer (or equivalently a group with homogeneous preferences as in Crawford and Meng): $\{\boldsymbol{p}_t, \boldsymbol{q}_t\}_{t=1,\ldots,T}$ with $\boldsymbol{p}_t \in \mathbb{R}^K_{++}$ and $\boldsymbol{q}_t \in \mathbb{R}^K_{+}$.

The constructions also involve a reference point $\boldsymbol{r}_t$ with $\boldsymbol{r}_t \in \mathbb{R}^K$ for each observation, taken (following Crawford and Meng) as a point expectation. (Sample variation ensures that realizations deviate from expectations.)

BCC focus on the leading case of multiple price-quantity observations for one consumer (or equivalently a group with homogeneous preferences as in Crawford and Meng): $\{\boldsymbol{p}_t, \boldsymbol{q}_t\}_{t=1,\dots,T}$ with $\boldsymbol{p}_t \in \mathbb{R}^K_{++}$ and $\boldsymbol{q}_t \in \mathbb{R}^K_{+}$.

The constructions also involve a reference point $\boldsymbol{r}_t$ with $\boldsymbol{r}_t \in \mathbb{R}^K$ for each observation, taken (following Crawford and Meng) as a point expectation. (Sample variation ensures that realizations deviate from expectations.)

Reference-dependent preferences are represented by a family of utility functions $u(\boldsymbol{q}, \boldsymbol{q} - \boldsymbol{r})$, indexed by the reference point (observable or not).

(This nests the standard reference-*in*dependent specification, where only levels matter; the case where only changes matter; and the Kőszegi and Rabin case where both levels and changes matter.)

BCC focus on the leading case of multiple price-quantity observations for one consumer (or equivalently a group with homogeneous preferences as in Crawford and Meng): $\{\boldsymbol{p}_t, \boldsymbol{q}_t\}_{t=1,\dots,T}$ with $\boldsymbol{p}_t \in \mathbb{R}^K_{++}$ and $\boldsymbol{q}_t \in \mathbb{R}^K_{+}$.

The constructions also involve a reference point $\boldsymbol{r}_t$ with $\boldsymbol{r}_t \in \mathbb{R}^K$ for each observation, taken (following Crawford and Meng) as a point expectation. (Sample variation ensures that realizations deviate from expectations.)

Reference-dependent preferences are represented by a family of utility functions $u(\boldsymbol{q}, \boldsymbol{q} - \boldsymbol{r})$, indexed by the reference point (observable or not).

(This nests the standard reference-*in*dependent specification, where only levels matter; the case where only changes matter; and the Kőszegi and Rabin case where both levels and changes matter.)

As in previous nonparametric demand analyses, BCC restrict attention to utility functions $u(\boldsymbol{q}, \boldsymbol{q} - \boldsymbol{r})$ that are continuous, non-satiated, and non-decreasing in consumption levels, and now changes.

($u(\boldsymbol{q}, \boldsymbol{q} - \boldsymbol{r})$ is just as flexible as an arbitrary continuous and non-satiated function $g(\boldsymbol{q}, \boldsymbol{r})$ would be, but easier to interpret. Continuity of $u(.)$ now includes continuity with respect to $\boldsymbol{r}$, a plausible restriction.)

Afriat (1967 *IER*) showed that a reference-*in*dependent utility function can rationalize the data if and only if the data satisfy GARP.

Definition 1. *(Rationalization). A reference-dependent utility function* $u(\boldsymbol{q}, \boldsymbol{q} - \boldsymbol{r})$ *and a set of reference points* $\{\boldsymbol{r}_t\}_{t=1,\ldots,T}$ *rationalize the data* $\{\boldsymbol{p}_t, \boldsymbol{q}_t\}_{t=1,\ldots,T}$ *iff*
$u(\boldsymbol{q}_t, \boldsymbol{q}_t - \boldsymbol{r}_t) \geq u(\boldsymbol{q}, \boldsymbol{q} - \boldsymbol{r}_t)$ *for all* $\boldsymbol{q}$ *such that* $\boldsymbol{p}_t^{'}\boldsymbol{q}_t \geq \boldsymbol{p}_t^{'}\boldsymbol{q}$.

Afriat (1967 *IER*) showed that a reference-*in*dependent utility function can rationalize the data if and only if the data satisfy GARP.

Definition 1. *(Rationalization). A reference-dependent utility function* $u(\boldsymbol{q}, \boldsymbol{q}-\boldsymbol{r})$ *and a set of reference points* $\{\boldsymbol{r}_t\}_{t=1,\dots,T}$ *rationalize the data* $\{\boldsymbol{p}_t, \boldsymbol{q}_t\}_{t=1,\dots,T}$ *iff*
$u(\boldsymbol{q}_t, \boldsymbol{q}_t-\boldsymbol{r}_t) \geq u(\boldsymbol{q}, \boldsymbol{q}-\boldsymbol{r}_t)$ *for all* $\boldsymbol{q}$ *such that* $\boldsymbol{p}_t' \boldsymbol{q}_t \geq \boldsymbol{p}_t' \boldsymbol{q}$.

Definition 2. *(Generalized Axiom of Revealed Preference; "GARP").* $\boldsymbol{q_s} R \boldsymbol{q_t}$ *implies* $\boldsymbol{p}_t'\boldsymbol{q}_t \leq \boldsymbol{p}_t'\boldsymbol{q}_s$ *where R indicates that there is some sequence of observations* $\boldsymbol{q}_i, \boldsymbol{q}_j, \boldsymbol{q}_k, \dots, \boldsymbol{q}_t$ *such that*
$\boldsymbol{p}_i'\boldsymbol{q}_i \geq \boldsymbol{p}_i'\boldsymbol{q}_j, \boldsymbol{p}_j'\boldsymbol{q}_j \geq \boldsymbol{p}_j'\boldsymbol{q}_k, \dots, \boldsymbol{p}_s'\boldsymbol{q}_s \geq \boldsymbol{p}_s'\boldsymbol{q}_t$.

Afriat (1967 *IER*) showed that a reference-*in*dependent utility function can rationalize the data if and only if the data satisfy GARP.

Definition 1. *(Rationalization). A reference-dependent utility function* $u(\boldsymbol{q}, \boldsymbol{q} - \boldsymbol{r})$ *and a set of reference points* $\{\boldsymbol{r}_t\}_{t=1,\dots,T}$ *rationalize the data* $\{\boldsymbol{p}_t, \boldsymbol{q}_t\}_{t=1,\dots,T}$ *iff*
$u(\boldsymbol{q}_t, \boldsymbol{q}_t - \boldsymbol{r}_t) \geq u(\boldsymbol{q}, \boldsymbol{q} - \boldsymbol{r}_t)$ *for all* $\boldsymbol{q}$ *such that* $\boldsymbol{p}_t^{'}\boldsymbol{q}_t \geq \boldsymbol{p}_t^{'}\boldsymbol{q}$.

Definition 2. *(Generalized Axiom of Revealed Preference; "GARP").* $\boldsymbol{q_s} R \boldsymbol{q_t}$ *implies* $\boldsymbol{p}_t^{'}\boldsymbol{q}_t \leq \boldsymbol{p}_t^{'}\boldsymbol{q}_s$ *where R indicates that there is some sequence of observations* $\boldsymbol{q}_i, \boldsymbol{q}_j, \boldsymbol{q}_k, \dots, \boldsymbol{q}_t$ *such that*
$\boldsymbol{p}_i^{'}\boldsymbol{q}_i \geq \boldsymbol{p}_i^{'}\boldsymbol{q}_j, \boldsymbol{p}_j^{'}\boldsymbol{q}_j \geq \boldsymbol{p}_j^{'}\boldsymbol{q}_k, \dots, \boldsymbol{p}_s^{'}\boldsymbol{q}_s \geq \boldsymbol{p}_s^{'}\boldsymbol{q}_t$.

We derive necessary and sufficient conditions for a reference-*dependent* rationalization for a sample with GARP violations, which that generalize those for a reference-*in*dependent rationalization.

Afriat (1967 *IER*) showed that a reference-*in*dependent utility function can rationalize the data if and only if the data satisfy GARP.

Definition 1. *(Rationalization). A reference-dependent utility function* $u(\boldsymbol{q}, \boldsymbol{q} - \boldsymbol{r})$ *and a set of reference points* $\{\boldsymbol{r}_t\}_{t=1,\ldots,T}$ *rationalize the data* $\{\boldsymbol{p}_t, \boldsymbol{q}_t\}_{t=1,\ldots,T}$ *iff*
$u(\boldsymbol{q}_t, \boldsymbol{q}_t - \boldsymbol{r}_t) \geq u(\boldsymbol{q}, \boldsymbol{q} - \boldsymbol{r}_t)$ *for all* $\boldsymbol{q}$ *such that* $\boldsymbol{p}_t^{'} \boldsymbol{q}_t \geq \boldsymbol{p}_t^{'} \boldsymbol{q}$.

Definition 2. *(Generalized Axiom of Revealed Preference; "GARP").* $\boldsymbol{q_s} R \boldsymbol{q_t}$ *implies* $\boldsymbol{p}_t^{'} \boldsymbol{q}_t \leq \boldsymbol{p}_t^{'} \boldsymbol{q}_s$ *where R indicates that there is some sequence of observations* $\boldsymbol{q}_i, \boldsymbol{q}_j, \boldsymbol{q}_k, \ldots, \boldsymbol{q}_t$ *such that*
$\boldsymbol{p}_i^{'} \boldsymbol{q}_i \geq \boldsymbol{p}_i^{'} \boldsymbol{q}_j, \boldsymbol{p}_j^{'} \boldsymbol{q}_j \geq \boldsymbol{p}_j^{'} \boldsymbol{q}_k, \ldots, \boldsymbol{p}_s^{'} \boldsymbol{q}_s \geq \boldsymbol{p}_s^{'} \boldsymbol{q}_t$.

We derive necessary and sufficient conditions for a reference-*dependent* rationalization for a sample with GARP violations, which that generalize those for a reference-*in*dependent rationalization.

Following Afriat, Diewert (1973 *RES*), and Varian (1982 *ECMA*), we give a tractable, linear programming method to check whether our conditions are satisfied and recover rationalizing preferences when they exist.

When reference points are unobservable, with diminishing sensitivity the indifference map can change with the reference point in an unrestricted way, making the hypothesis of rationality with reference-dependent preferences irrefutable:

When reference points are unobservable, with diminishing sensitivity the indifference map can change with the reference point in an unrestricted way, making the hypothesis of rationality with reference-dependent preferences irrefutable:

Proposition 1. *(Diminishing sensitivity with unobservable reference points). For any dataset* $\{\boldsymbol{p}_t, \boldsymbol{q}_t\}_{t=1,\dots,T}$*, there exists a set of reference points* $\{\boldsymbol{r}_t\}_{t=1,\dots,T}$ *and a reference-dependent utility function* $u(\boldsymbol{q}, \boldsymbol{q} - \boldsymbol{r})$ *which is continuous, nonsatiated, and non-decreasing with respect to* $\boldsymbol{q}$ *for a given* $\boldsymbol{r}$ *which rationalizes those data.*

When reference points are unobservable, with diminishing sensitivity the indifference map can change with the reference point in an unrestricted way, making the hypothesis of rationality with reference-dependent preferences irrefutable:

Proposition 1. *(Diminishing sensitivity with unobservable reference points). For any dataset* $\{\boldsymbol{p}_t, \boldsymbol{q}_t\}_{t=1,\ldots,T}$*, there exists a set of reference points* $\{\boldsymbol{r}_t\}_{t=1,\ldots,T}$ *and a reference-dependent utility function* $u(\boldsymbol{q}, \boldsymbol{q}-\boldsymbol{r})$ *which is continuous, nonsatiated, and non-decreasing with respect to* $\boldsymbol{q}$ *for a given* $\boldsymbol{r}$ *which rationalizes those data.*

Proof. We can use reference-dependence to make the indifference curve through each observation coincide with its budget line. Let $u(\boldsymbol{q}, \boldsymbol{q}-\boldsymbol{r}) = \boldsymbol{r}'\boldsymbol{q} + \boldsymbol{r}'(\boldsymbol{q}-\boldsymbol{r})$ which is continuous, nonsatiated and non-decreasing with respect to $\boldsymbol{q}$ for a given $\boldsymbol{r} > 0$. Set $\boldsymbol{r}_t = \boldsymbol{p}_t$. (Although $\boldsymbol{r}_t$ is normally conformable to $\boldsymbol{q}_t$, not $\boldsymbol{p}_t$, $\boldsymbol{p}_t$ serves here only to identify the marginal utilities of the approximating linear preferences.) Then $u(\boldsymbol{q}_t, \boldsymbol{q}_t-\boldsymbol{r}_t) = 2\boldsymbol{p}_t'\boldsymbol{q}_t - \boldsymbol{p}_t'\boldsymbol{p}_t$ and $u(\boldsymbol{q}, \boldsymbol{q}-\boldsymbol{r}_t) = 2\boldsymbol{p}_t'\boldsymbol{q} - \boldsymbol{p}_t'\boldsymbol{p}_t$.■

Constant sensitivity is a useful, if oversimplified, way to rule out the theoretically possible but empirically implausible strong local variations in preferences that prevent the model from having testable implications with diminishing sensitivity.

Constant sensitivity is a useful, if oversimplified, way to rule out the theoretically possible but empirically implausible strong local variations in preferences that prevent the model from having testable implications with diminishing sensitivity.

A reference point (observed or hypothesized) partitions the commodity space into 2^K *reference regimes*. Observations are in the same regime (whether or not they have exactly the same reference point) if and only if they have the same good-by-good gain-loss pattern. Let $\mathrm{sign}(\boldsymbol{q} - \boldsymbol{r})$ denote the vector whose *k*-th component is $\mathrm{sign}(q_k - r_k)$.

Constant sensitivity is a useful, if oversimplified, way to rule out the theoretically possible but empirically implausible strong local variations in preferences that prevent the model from having testable implications with diminishing sensitivity.

A reference point (observed or hypothesized) partitions the commodity space into 2^K *reference regimes*. Observations are in the same regime (whether or not they have exactly the same reference point) if and only if they have the same good-by-good gain-loss pattern. Let $\mathrm{sign}(\boldsymbol{q} - \boldsymbol{r})$ denote the vector whose *k*-th component is $\mathrm{sign}(q_k - r_k)$.

Definition 3. *(Constant Sensitivity). A reference-dependent utility function* $u(\boldsymbol{q}, \boldsymbol{q} - \boldsymbol{r})$ *exhibits constant sensitivity/sign-dependence if for any two bundles* $\boldsymbol{q}^*$ *and* $\boldsymbol{q}$ *it is the case that* $u(\boldsymbol{q}, \boldsymbol{q} - \boldsymbol{r}) \geq u(\boldsymbol{q}^*, \boldsymbol{q}^* - \boldsymbol{r}) \Leftrightarrow u(\boldsymbol{q}, \boldsymbol{q} - \boldsymbol{r}^*) \geq u(\boldsymbol{q}^*, \boldsymbol{q}^* - \boldsymbol{r}^*)$ *if* $\mathrm{sign}(\boldsymbol{q} - \boldsymbol{r}) = \mathrm{sign}(\boldsymbol{q}^* - \boldsymbol{r}) = \mathrm{sign}(\boldsymbol{q} - \boldsymbol{r}^*) = \mathrm{sign}(\boldsymbol{q}^* - \boldsymbol{r}^*)$.

Under constant sensitivity each gain-loss regime has its own *local* indifference map, which extends throughout the space but is “switched on” only for consumption bundles in that regime.

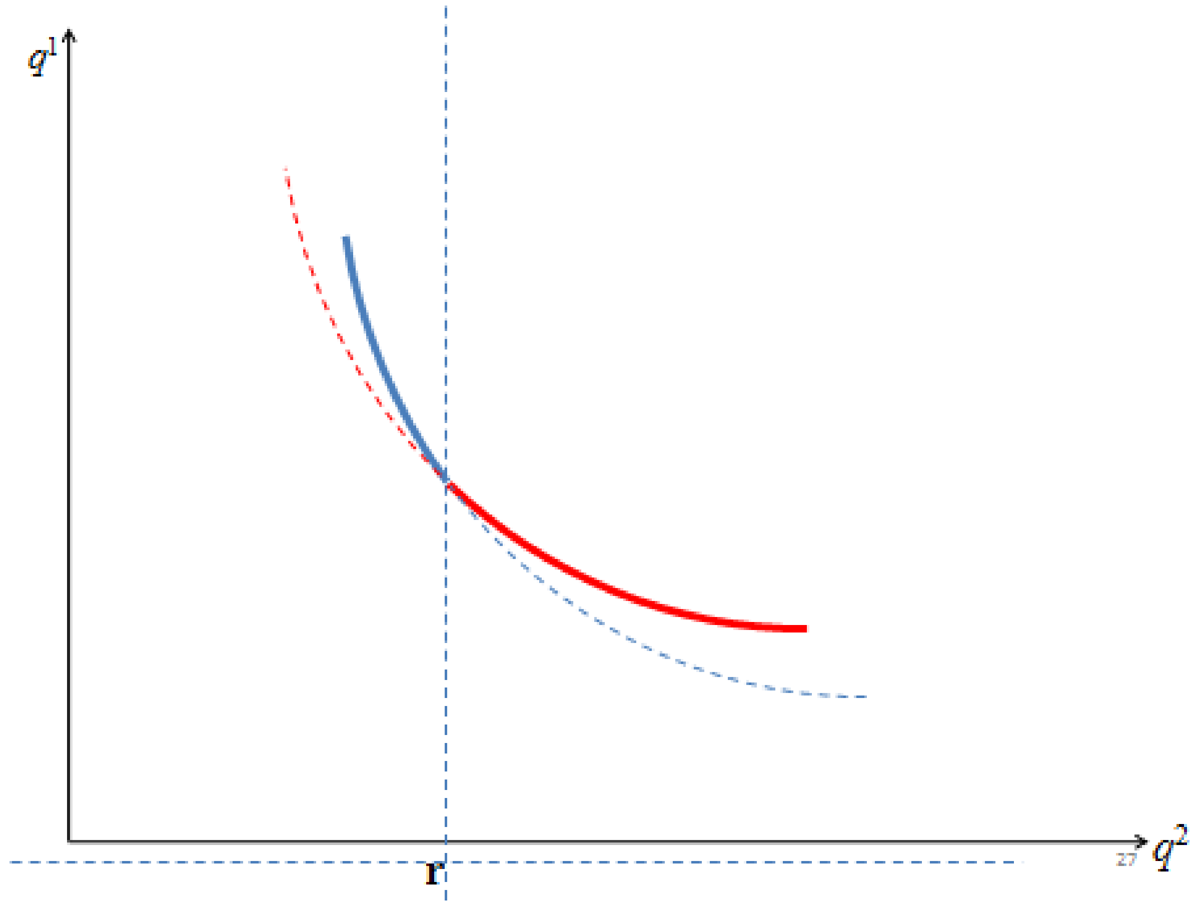
Under constant sensitivity each gain-loss regime has its own *local* indifference map, which extends throughout the space but is “switched on” only for consumption bundles in that regime.

Although constant sensitivity requires the local map to remain constant within a regime, it allows the *level* of $u(\boldsymbol{q}, \boldsymbol{q} - \boldsymbol{r})$ to vary with $(\boldsymbol{q} - \boldsymbol{r})$.

Under constant sensitivity each gain-loss regime has its own *local* indifference map, which extends throughout the space but is “switched on” only for consumption bundles in that regime.

Although constant sensitivity requires the local map to remain constant within a regime, it allows the *level* of $u(\boldsymbol{q}, \boldsymbol{q} - \boldsymbol{r})$ to vary with $(\boldsymbol{q} - \boldsymbol{r})$.

We assume that preferences are continuous and weakly monotonic, so that for any $\boldsymbol{r}$ the local maps join uniquely across regimes to create a well-defined *global* indifference map.



Constant sensitivity with two active gain-loss regimes

Under constant sensitivity each gain-loss regime has its own *local* indifference map, which extends throughout the space but is "switched on" only for consumption bundles in that regime.

Although constant sensitivity requires the local map to remain constant within a regime, it allows the *level* of $u(\boldsymbol{q}, \boldsymbol{q} - \boldsymbol{r})$ to vary with $(\boldsymbol{q} - \boldsymbol{r})$.

We assume that preferences are continuous and weakly monotonic, so that for any $\boldsymbol{r}$ the local maps join uniquely across regimes to create a well-defined *global* indifference map.

Although the local indifference maps are insensitive to the precise location of $\boldsymbol{r}$, even small changes in $\boldsymbol{r}$ vary how they connect across regimes, altering the shape of the global map.

When reference points are unobservable, with constant sensitivity choice can be rationalized by reference-dependent preferences if and only if one can hypothesize reference points that group the observations into 2^K gain-loss regimes (where K is the number of goods), such that each regime's observations satisfy GARP, so that they are rationalizable within their regime by reference-*in*dependent preferences:

When reference points are unobservable, with constant sensitivity choice can be rationalized by reference-dependent preferences if and only if one can hypothesize reference points that group the observations into 2^K gain-loss regimes (where K is the number of goods), such that each regime's observations satisfy GARP, so that they are rationalizable within their regime by reference-*in*dependent preferences:

Proposition 2. *(Constant sensitivity with unobservable reference points). The following conditions are equivalent:*

1. There exist a set of reference points $\{\boldsymbol{r}_t\}_{t=1,\ldots,T}$ *and a reference-dependent utility function* $u(\boldsymbol{q}, \boldsymbol{q}-\boldsymbol{r})$ *which exhibits constant sensitivity and which is continuous, nonsatiated, and non-decreasing with respect to* $\boldsymbol{q}$ *for a given* $\boldsymbol{r}$ *which rationalize the data* $\{\boldsymbol{p}_t, \boldsymbol{q}_t\}_{t=1,\ldots,T}$.

2. There exists an exclusive and exhaustive partition of the data $\{\boldsymbol{p}_t, \boldsymbol{q}_t\}_{t=1,\ldots,T}$ *into* 2^K *or fewer subsets such that GARP is satisfied within each subset.*

Sketch of proof.

$(1) \Rightarrow (2)$: Plainly if we cannot satisfy GARP within some specification of 2^K regimes, we cannot rationalize the data.

$(2) \Rightarrow (1)$: Suppose that there exists a partition of the observations into 2^K subsets within which the data satisfy GARP. The cross-regime restrictions for rationalization can always be satisfied by hypothesizing reference points that put each observation's entire budget set within its reference regime. ■

When reference points are observable, Proposition 1's result that with diminishing sensitivity the hypothesis of reference-dependent preferences is irrefutable, remains valid with a minor qualification.

Proposition 3. *(Diminishing sensitivity with observable reference points). The following conditions are equivalent:*

1. There exists a reference-dependent utility function $u(\boldsymbol{q}, \boldsymbol{q} - \boldsymbol{r})$ *which is continuous, nonsatiated and non-decreasing with respect to* $\boldsymbol{q}$ *which rationalizes the data* $\{\boldsymbol{p}_t, \boldsymbol{q}_t, \boldsymbol{r}_t\}_{t=1,\ldots,T}$.

2. Every subset of the data which is defined by having a common reference point satisfies GARP.

Proof of Proposition 3.

(1) ⇒ (2): For a fixed reference point, $u(\boldsymbol{q}, \boldsymbol{q} - \boldsymbol{r})$ is continuous, nonsatiated and non-decreasing with respect to $\boldsymbol{q}$ and therefore implies GARP by Afriat's Theorem.

(2) ⇒ (1): In any subset with a common reference point, GARP and Afriat's Theorem implies that, as in the proof of Proposition 2, we can find a utility function that rationalizes the data for that subset. For other observations we can use reference-dependence to construct such a utility function (but no longer as immediately as in the proof of Proposition 1).

Corollary 1. *(Reference dependence with a fixed reference point). The following conditions are equivalent:*

1. There exists a reference-dependent utility function $u(\boldsymbol{q}, \boldsymbol{q} - \boldsymbol{r})$ *which is continuous, nonsatiated and non-decreasing with respect to* $\boldsymbol{q}$ *which rationalizes the dataset* $\{\boldsymbol{p}_t, \boldsymbol{q}_t, \boldsymbol{r}_t\}_{t=1,\dots,T}$ *where* $\boldsymbol{r}_t = \boldsymbol{r}$ *for all* t.

2. There exists a utility function $u(\boldsymbol{q})$ *which is continuous, concave, nonsatiated and non-decreasing with respect to* $\boldsymbol{q}$ *which rationalizes the dataset* $\{\boldsymbol{p}_t, \boldsymbol{q}_t\}_{t=1,\dots,T}$

3. The data $\{\boldsymbol{p}_t, \boldsymbol{q}_t\}_{t=1,\dots,T}$ *satisfies GARP.*

Proof of Corollary 1.

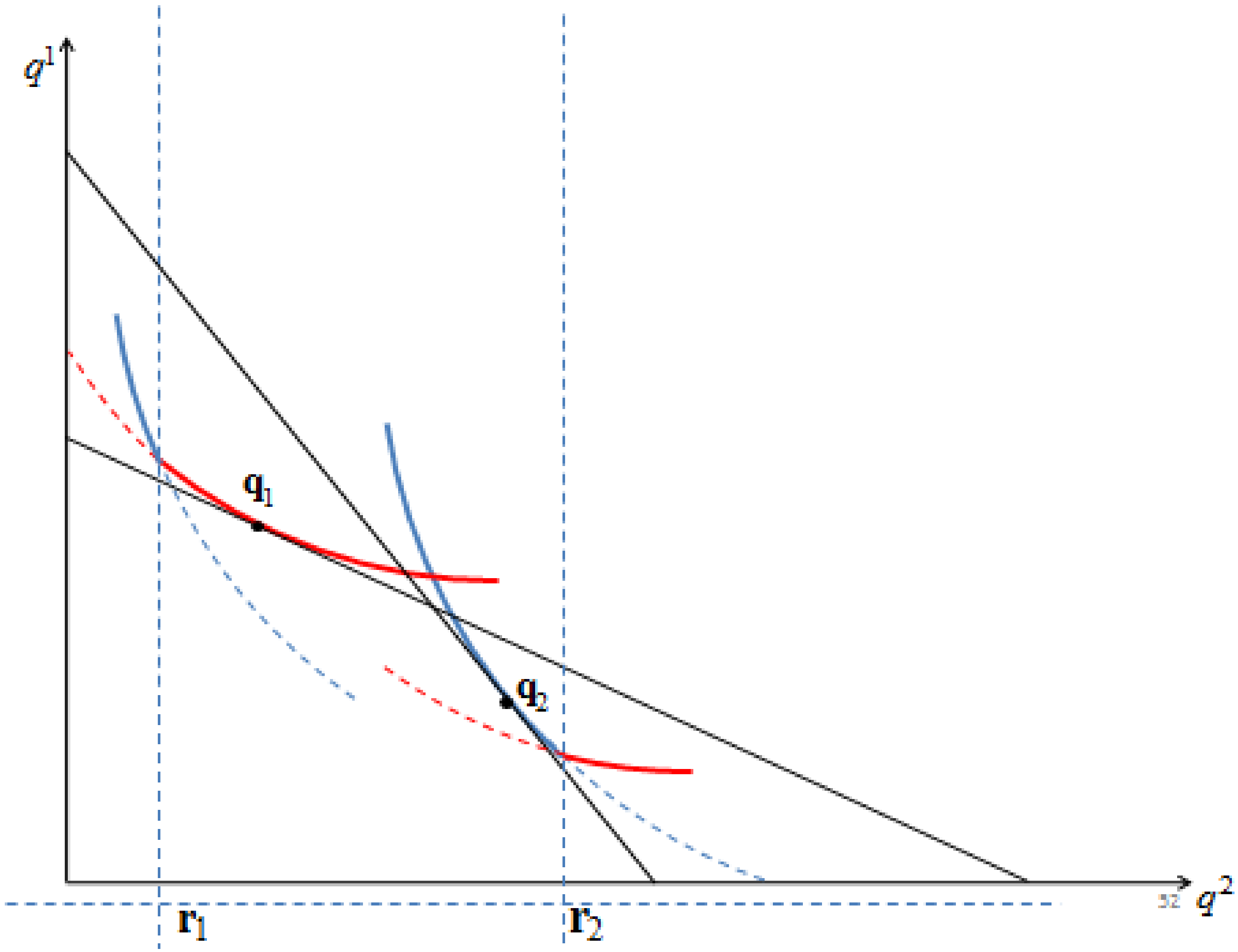
With all observations sharing the same, single reference point, this follows from Proposition 2. ■

Proposition 4. *(Constant sensitivity with observable reference points). The following conditions are equivalent:*

1. There exists a reference-dependent utility function $u(\boldsymbol{q}, \boldsymbol{q} - \boldsymbol{r})$ *which exhibits constant sensitivity and which is continuous, nonsatiated and non-decreasing with respect to* $\boldsymbol{q}$ *for a given* $\boldsymbol{r}$ *which rationalizes the data* $\{\boldsymbol{p}_t, \boldsymbol{q}_t, \boldsymbol{r}_t\}_{t=1,\ldots,T}$.

2. The data within each regime defined by the reference points satisfy GARP and the implied revealed preferred and revealed worse sets (as defined by Varian but not explained here) for each observation are disjoint.

Proof omitted.



Rationalizing a cross-regime GARP violation with constant sensitivity

Application to cabdrivers' labor supply

BCC use their methods to reconsider Farber's (2005, 2008) and Crawford and Meng's (2011) analyses of Farber's dataset from a nonparametric point of view.

Application to cabdrivers' labor supply

BCC use their methods to reconsider Farber's (2005, 2008) and Crawford and Meng's (2011) analyses of Farber's dataset from a nonparametric point of view. The model of interest is

$$\max_{\boldsymbol{q}^i} u^i(\boldsymbol{q}^i, \boldsymbol{q}^i - \boldsymbol{r}_t^i) \text{ subject to } \boldsymbol{p}_t^{i\prime}\boldsymbol{q}^i = x_t^i$$

where x_t^i is the value of their time endowment at the hourly wage w_t^i.

Application to cabdrivers' labor supply

BCC use their methods to reconsider Farber's (2005, 2008) and Crawford and Meng's (2011) analyses of Farber's dataset from a nonparametric point of view. The model of interest is

$$\max_{\boldsymbol{q}^i} u^i(\boldsymbol{q}^i, \boldsymbol{q}^i - \boldsymbol{r}_t^i) \text{ subject to } \boldsymbol{p}_t^{i\prime}\boldsymbol{q}^i = x_t^i$$

where x_t^i is the value of their time endowment at the hourly wage w_t^i.

Given Propositions 1's and 3's negative results, BCC assume constant sensitivity throughout.

Application to cabdrivers' labor supply

BCC use their methods to reconsider Farber's (2005, 2008) and Crawford and Meng's (2011) analyses of Farber's dataset from a nonparametric point of view. The model of interest is

$$\max_{\boldsymbol{q}^i} u^i(\boldsymbol{q}^i, \boldsymbol{q}^i - \boldsymbol{r}_t^i) \text{ subject to } \boldsymbol{p}_t^i{}'\boldsymbol{q}^i = x_t^i$$

where x_t^i is the value of their time endowment at the hourly wage w_t^i.

Given Propositions 1's and 3's negative results, BCC assume constant sensitivity throughout.

Partly for computational feasibility, BCC follow the nonparameteric tradition of modeling each driver separately, by contrast with the labor-economics tradition of assuming drivers all have the same preferences.

("Point" estimates are feasible, possibly even for latent class models. The problem is computing the Selten measures of goodness of fit.)

Farber collected 538 “trip sheets” for 15 drivers between June 1999 and May 2001. Each trip sheet records the driver's name, hack number, and date and the details of each fare. For each fare the data record the start time, start location, end time, end location and fare.

Farber collected 538 “trip sheets” for 15 drivers between June 1999 and May 2001. Each trip sheet records the driver's name, hack number, and date and the details of each fare. For each fare the data record the start time, start location, end time, end location and fare.

Create a price-quantity panel dataset $\{\boldsymbol{p}_t^i, \boldsymbol{q}_t^i\}_{t=1,\ldots,T^i}^{i=1,\ldots,15}$ where $i = 1, \ldots, 15$ indicates a driver and $t = 1, \ldots, T^i$ indexes observations (shifts) for each driver. The first shift for each driver is $t = 1$ and the total number of shifts observed for each driver is T_i and is different for each driver.

Farber collected 538 “trip sheets” for 15 drivers between June 1999 and May 2001. Each trip sheet records the driver's name, hack number, and date and the details of each fare. For each fare the data record the start time, start location, end time, end location and fare.

Create a price-quantity panel dataset $\{\boldsymbol{p}_t^i, \boldsymbol{q}_t^i\}_{t=1,\ldots,T^i}^{i=1,\ldots,15}$ where $i = 1, \ldots, 15$ indicates a driver and $t = 1, \ldots, T^i$ indexes observations (shifts) for each driver. The first shift for each driver is $t = 1$ and the total number of shifts observed for each driver is T_i and is different for each driver.

BCC define two choice variables, leisure time $\left(l_t^i\right)$ and consumption $\left(y_t^i\right)$ and their corresponding prices:

$$\boldsymbol{q}_t^i = \begin{bmatrix} l_t^i \\ y_t^i \end{bmatrix} = \begin{bmatrix} 24 - \textit{work hours} - \textit{breaks} - \textit{waiting} \\ \textit{earnings} \end{bmatrix}$$

$$\boldsymbol{p}_t^i = \begin{bmatrix} w_t^i \\ 1 \end{bmatrix} = \begin{bmatrix} \textit{hourly earnings} \\ \textit{numeraire} \end{bmatrix}$$

where hourly earnings is adjusted for waiting (and not earning) time.

BCC begin by checking whether any of the drivers satisfy the weak axiom of revealed preference (“WARP”): that is, whether their choices can be rationalized by standard, reference-*in*dependent preferences.

BCC present two measures of the results of this and later tests. The first is the “pass rate”, $r^i \in [0,1]$, which does not reflect the model’s flexibility.

BCC begin by checking whether any of the drivers satisfy the weak axiom of revealed preference ("WARP"): that is, whether their choices can be rationalized by standard, reference-*in*dependent preferences.

BCC present two measures of the results of this and later tests. The first is the "pass rate", $r^i \in [0,1]$, which does not reflect the model's flexibility.

The second, $m^i = r^i - a^i \in [-1,1]$, is the difference between r^i and the "area" a^i, the size of the set of all possible choices (Selten and Krischker 1983).

The second measure rewards a model for a good pass rate despite demanding restrictions, essential comparing models that vary in flexibility:

An unrestrictive model yields $m^i = 0$, a pass despite sharp restrictions yields $m^i = 1$, and a failure despite weak restrictions yields $m^i = -1$.

Unobserved reference points

With constant sensitivity, when reference points are unobserved, Proposition 2 shows that the test for whether a driver's choices are consistent with reference-dependent preferences reduces to determining how many indifference maps are required to rationalize them.

Unobserved reference points

With constant sensitivity, when reference points are unobserved, Proposition 2 shows that the test for whether a driver's choices are consistent with reference-dependent preferences reduces to determining how many indifference maps are required to rationalize them.

With only two goods, WARP is necessary and sufficient for utility maximization. Thus if a driver's observations satisfy WARP, perfect fit can be achieved with a single, reference-*in*dependent indifference map.

Unobserved reference points

With constant sensitivity, when reference points are unobserved, Proposition 2 shows that the test for whether a driver's choices are consistent with reference-dependent preferences reduces to determining how many indifference maps are required to rationalize them.

With only two goods, WARP is necessary and sufficient for utility maximization. Thus if a driver's observations satisfy WARP, perfect fit can be achieved with a single, reference-*in*dependent indifference map.

If not, BCC check whether a driver's observations can be partitioned into two subsets, within each of which WARP is satisfied and they can be rationalized by a single indifference map. If so, the observations can be rationalized by reference-dependent preferences with a target for hours *or* earnings (which, is not identified with unobservable reference points).

Unobserved reference points

With constant sensitivity, when reference points are unobserved, Proposition 2 shows that the test for whether a driver's choices are consistent with reference-dependent preferences reduces to determining how many indifference maps are required to rationalize them.

With only two goods, WARP is necessary and sufficient for utility maximization. Thus if a driver's observations satisfy WARP, perfect fit can be achieved with a single, reference-*in*dependent indifference map.

If not, BCC check whether a driver's observations can be partitioned into two subsets, within each of which WARP is satisfied and they can be rationalized by a single indifference map. If so, the observations can be rationalized by reference-dependent preferences with a target for hours *or* earnings (which is not identified with unobservable reference points).

If two maps do not suffice, BCC check whether a driver's observations can be partitioned into four (= 2^2) subsets, within each of which WARP is satisfied. If so, the observations can be rationalized by reference-dependent preferences with targets for both hours and earnings.

The cross-driver average pass rates for the neoclassical, two-map, and four-map models are 86.36%, 99.47% and 100% respectively.

Is the 13-14% increase in from neoclassical to reference-dependent models enough to justify the latter models' extra flexibility?

The cross-driver average pass rates for the neoclassical, two-map, and four-map models are 86.36%, 99.47% and 100% respectively.

Is the 13-14% increase in from neoclassical to reference-dependent models enough to justify the latter models' extra flexibility?

The four-map (two-target reference-dependent) model does poorly for all 15 drivers, hardly better than random, with a very high pass rate but also, with unobserved reference points and a small sample, a very high area.

The cross-driver average pass rates for the neoclassical, two-map, and four-map models are 86.36%, 99.47% and 100% respectively.

Is the 13-14% increase in from neoclassical to reference-dependent models enough to justify the latter models' extra flexibility?

The four-map (two-target reference-dependent) model does poorly for all 15 drivers, hardly better than random, with a very high pass rate but also, with unobserved reference points and a small sample, a very high area.

The two-map (one-target reference-dependent) model does best for 7 drivers, and the one-map (reference-*in*dependent) model does best for 8.

(The cross-driver heterogeneity in the Selten indices is due to drivers' different numbers of observations and the variability of hourly wages.)

The cross-driver average pass rates for the neoclassical, two-map, and four-map models are 86.36%, 99.47% and 100% respectively.

Is the 13-14% increase in from neoclassical to reference-dependent models enough to justify the latter models' extra flexibility?

The four-map (two-target reference-dependent) model does poorly for all 15 drivers, hardly better than random, with a very high pass rate but also, with unobserved reference points and a small sample, a very high area.

The two-map (one-target reference-dependent) model does best for 7 drivers, and the one-map (reference-*in*dependent) model does best for 8.

(The cross-driver heterogeneity in the Selten indices is due to drivers' different numbers of observations and the variability of hourly wages.)

Overall, with unobserved reference points the reference-*in*dependent model does best: Its area is much smaller than those of both reference-dependent models, but its pass rate is still high, at 86.36%.

Table 2: Selten indices with unobserved reference points

Driver	One map	Two Maps	Four Maps
1	0.8718	0.8730	0.0000
2	0.9360	0.2050	0.0000
3	0.9250	0.8240	0.0010
4	0.7826	0.4100	0.0000
5	0.9583	0.5920	0.0000
6	0.8919	0.9160	0.0050
7	0.8321	0.3100	0.0010
8	0.8889	0.9380	0.0010
9	0.8811	0.2120	0.0000
10	0.7627	0.3880	0.0000
11	0.8714	0.9847	0.0150
12	0.7639	0.9861	0.0450
13	0.8182	0.6797	0.0000
14	0.8261	0.9593	0.0080
15	0.8261	0.9570	0.0060

Observed reference points

With constant sensitivity, when reference points are observed or as here, modeled as functions of the data, Proposition 4 is the relevant result.

BCC follow Koszegi and Rabin (2006) in conceptualizing the reference points as rational expectations; and Crawford and Meng (2011) in treating reference points as point expectations, driver by driver.

BCC model them from a driver's observations in two alternative ways:

- Proxy by leave-one-out means (close to Crawford and Meng 2011).

- Proxy by a backward-looking, lagged model in which the reference point depends on what the driver chose on his last comparable shift.

BCC cross the leave-one-out mean and lagged reference point models with distinguishing day from night shifts and rainy from dry shifts.

(E.g. the reference point for a rainy shift with the backward-looking model is what happened on the driver's last comparably *rainy* shift.)

Altogether there are 18 alternative reference-point models.

Table 3 reports Selten indices, driver by driver, for the models with observed/modeled reference points with respect to hours.

If all drivers are assumed to form their reference points in the same way, all 18 models have reasonable Selten indices (0.7246 on average), with cross-driver averages ranging only from 0.7047 to 0.7627.

Table 3 reports Selten indices, driver by driver, for the models with observed/modeled reference points with respect to hours.

If all drivers are assumed to form their reference points in the same way, all 18 models have reasonable Selten indices (0.7246 on average), with cross-driver averages ranging only from 0.7047 to 0.7627.

The best model is the lagged conditional on weather one (the average Selten index across drivers is 0.7627. with median 0.7857).

Table 3 reports Selten indices, driver by driver, for the models with observed/modeled reference points with respect to hours.

If all drivers are assumed to form their reference points in the same way, all 18 models have reasonable Selten indices (0.7246 on average), with cross-driver averages ranging only from 0.7047 to 0.7627.

The best model is the lagged conditional on weather one (the average Selten index across drivers is 0.7627. with median 0.7857).

If drivers are allowed to form their reference points heterogeneously, the day/night lagged model is best (or joint best) for 8 drivers; the unconditional leave-one-out model is best for driver 4; and the leave-one-out day/night model is best for drivers 8 and 9.

Table 3 reports Selten indices, driver by driver, for the models with observed/modeled reference points with respect to hours.

If all drivers are assumed to form their reference points in the same way, all 18 models have reasonable Selten indices (0.7246 on average), with cross-driver averages ranging only from 0.7047 to 0.7627.

The best model is the lagged conditional on weather one (the average Selten index across drivers is 0.7627. with median 0.7857).

If drivers are allowed to form their reference points heterogeneously, the day/night lagged model is best (or joint best) for 8 drivers; the unconditional leave-one-out model is best for driver 4; and the leave-one-out day/night model is best for drivers 8 and 9.

Overall, allowing heterogeneity in how drivers form reference points and choosing the best model for each driver, the average Selten index for observed reference points with respect to hours is 0.8122.

Table 3: Selten indices with observed reference points with respect to hours

		Leave-one-out			Lagged	
Driver		**Day/Night**	**Rain/Dry**		**Day/Night**	**Rain/Dry**
1	0.7692	0.7436	0.7692	0.8462	0.7692	0.8462
2	0.7430	0.7730	0.7190	0.8290	0.8500	0.8130
3	0.8000	0.7480	0.7990	0.7250	0.9250	0.7240
4	0.6537	0.6082	0.6507	0.5552	0.5197	0.5582
5	0.8550	0.8610	0.8610	0.8333	0.9553	0.8303
6	0.7297	0.6757	0.7277	0.8919	0.8919	0.8108
7	0.6062	0.5992	0.6418	0.8657	0.8657	0.8647
8	0.8444	0.8667	0.7323	0.7111	0.7333	0.6889
9	0.7350	0.7570	0.7230	0.5034	0.6093	0.4984
10	0.5591	0.7615	0.6309	0.7517	0.8714	0.6401
11	0.8000	0.7429	0.8000	0.8286	0.7857	0.8286
12	0.6806	0.6806	0.4583	0.6806	0.6250	0.6806
13	0.8162	0.8182	0.8182	0.7273	0.9091	0.7263
14	0.6304	0.5217	0.6304	0.5217	0.4783	0.5217
15	0.6087	0.6087	0.6087	0.6522	0.6522	0.6522

Table 4 reports Selten indices, driver by driver, for the models with observed reference points with respect to earnings.

This time, if all drivers are assumed to form their reference points in the same way, the day/night leave-one-out and lagged models are the best, with average Selten indices of 0.7501 and 0.7426 respectively.

Allowing heterogeneity in how drivers form their reference points, looking across drivers the day/night lagged model is the best for 6 drivers.

Overall, allowing heterogeneity in how drivers form reference points and choosing the best model for each driver, the average Selten index for observed reference points with respect to hours is 0.8122.

Table 4: Selten indices with observed reference points with respect to earnings

	Leave-one-out			Lagged		
Driver		Day/Night	Rain/Dry		Day/Night	Rain/Dry
1	0.8462	0.8205	0.8462	0.7179	0.7436	0.7179
2	0.7350	0.7500	0.7330	0.8350	0.8460	0.8360
3	0.8250	0.7730	0.8250	0.6750	0.9000	0.7000
4	0.9275	0.9235	0.9145	0.7706	0.9445	0.7766
5	0.8947	0.9413	0.8987	0.8323	0.9563	0.8293
6	0.5946	0.5946	0.5666	0.7297	0.6757	0.7027
7	0.7175	0.8097	0.5782	0.9004	0.6462	0.9340
8	0.6667	0.6444	0.7111	0.7333	0.6667	0.7556
9	0.7150	0.7510	0.7000	0.5034	0.4774	0.4964
10	0.7135	0.7535	0.6867	0.8095	0.7617	0.7497
11	0.7286	0.6429	0.7286	0.6571	0.7857	0.6571
12	0.6528	0.6806	0.5694	0.5833	0.6389	0.5972
13	0.8162	0.8182	0.8182	0.6970	0.8788	0.6647
14	0.6957	0.6522	0.6957	0.4783	0.5870	0.4783
15	0.6957	0.6957	0.6957	0.6087	0.6304	0.6087

Table 5 reports Selten indices, driver by driver, for the models with observed reference points with respect to both hours and earnings.

If all drivers are assumed to form their reference points in the same way, the day/night lagged model is best, with average Selten index of 0.7651.

Allowing heterogeneity in how drivers form their reference points, looking across drivers the day/night lagged model is the best for 10 drivers.

Overall, allowing heterogeneity in how drivers form reference points, the average Selten index for observed reference points with respect to both earnings and hours is 0.7028, versus 0.7291 for earnings only and 0.7246 for hours only.

And choosing the best both-hours-and-earnings model for each driver, the average Selten index for observed reference points is 0.8069, versus 0.8139 for earnings only and 0.8122 for hours only.

Table 5: Selten indices for observed reference points with respect to both earnings and hours

	Leave-one-out			Lagged		
Driver		**Day/Night**	**Rain/Dry**		**Day/Night**	**Rain/Dry**
1	0.7939	0.7949	0.7949	0.7179	0.7949	0.7179
2	0.6980	0.6850	0.6580	0.7280	0.7340	0.6530
3	0.7000	0.6710	0.7230	0.6750	0.8990	0.6990
4	0.8895	0.8865	0.8785	0.7596	0.9305	0.7626
5	0.8043	0.9273	0.8043	0.8293	0.9483	0.8263
6	0.5135	0.4595	0.6727	0.8919	0.9459	0.8649
7	0.5872	0.7867	0.5652	0.8177	0.8067	0.9130
8	0.6444	0.6222	0.5990	0.7111	0.6222	0.7111
9	0.6880	0.7100	0.6810	0.3625	0.3215	0.3485
10	0.6467	0.7255	0.7245	0.7885	0.8464	0.7785
11	0.6857	0.6429	0.6857	0.8286	0.8429	0.8286
12	0.6389	0.6389	0.5833	0.6111	0.7222	0.6389
13	0.7516	0.7566	0.7566	0.6960	0.9091	0.7253
14	0.5000	0.4783	0.5000	0.4565	0.5217	0.4565
15	0.5870	0.6739	0.5870	0.5870	0.6304	0.5870

Summing up, reference-dependence allows a parsimonious nonparametric rationalization of most drivers’ labor supply decisions, and identifies the key elements of such a rationalization.

Comparing models’ Selten indices, the neoclassical reference-*in*dependent model is best only 4 drivers, while some form of reference-dependence does best for 11 drivers.

Among the alternative reference-dependent models, ones with one target two-regimes but reference unspecified is best for 6 drivers.

For the other drivers, the day/night lagged model is best.