The 22nd Behavior Modeling Summer School

Sep. 18 – 20 , 2023 @ The University of Tokyo

Size Matters:

The Use and Misuse of Statistical Significance in Discrete Choice Models

Giancarlos Parady – The University of Tokyo

Parady, G., Ory, D., Walker, J. (2021) The overreliance on statistical goodness of fit and under-reliance on validation in discrete choice models: A review of validation practices in the transportation academic literature . Journal of Choice Modelling 38, 100257 (Open Access)

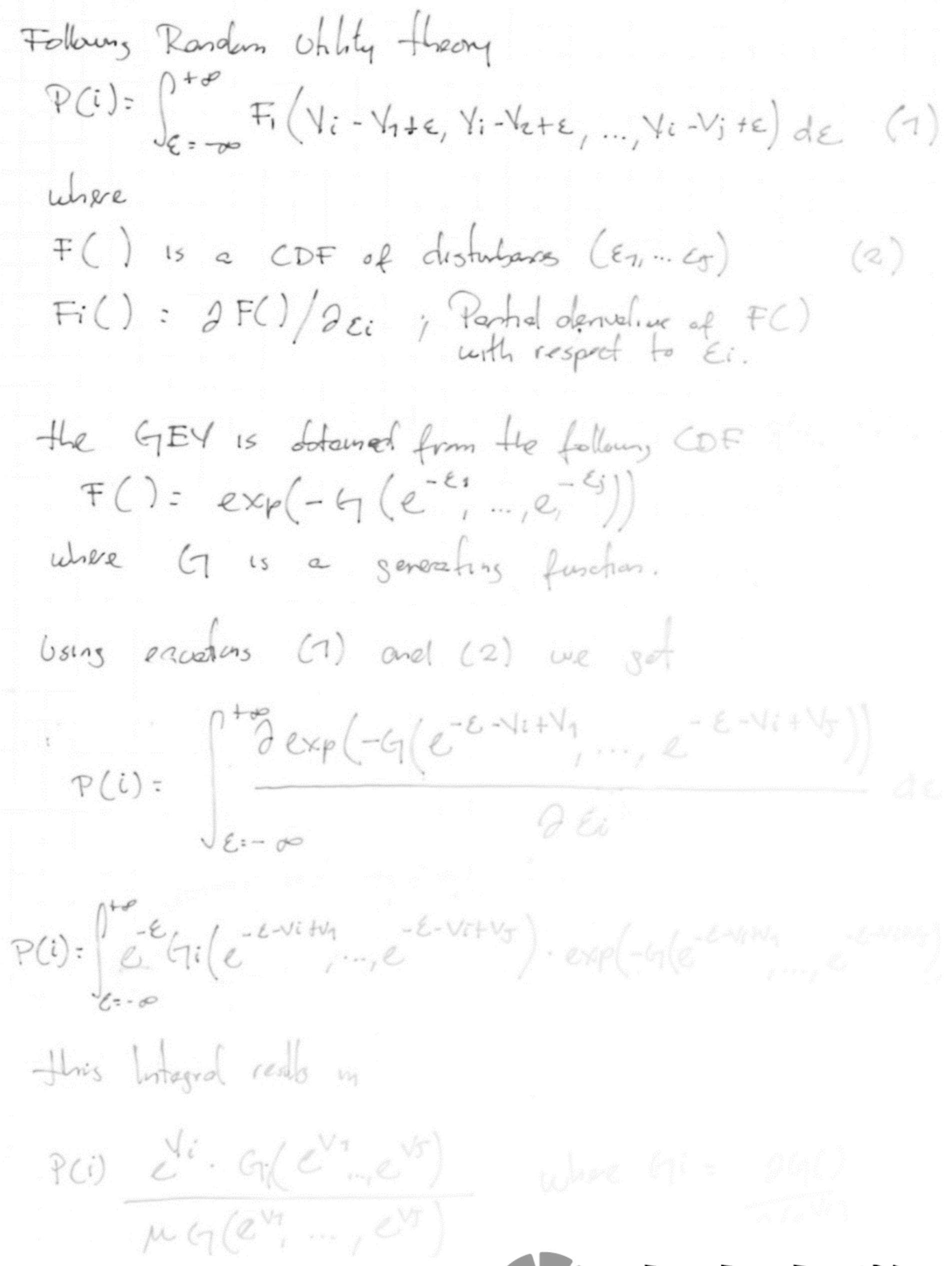
Parady, G., Axhausen K.W. (2023) Size Matters: The Use and Misuse of Statistical Significance in Discrete Choice Models in the Transportation Academic Literature. Transportation (accepted)

東京大学 THE UNIVERSITY OF TOKYO

Size Matters:

The Use and Misuse of Statistical Significance in Discrete Choice Models

Basic inference with discrete choice models



東京大学
THE UNIVERSITY OF TOKYO

Why is inference important?

Variable name	Coefficient	S.E.	t statistic
Auto constant	1.45	0.393	3.70
In-vehicle time (min)	-0.0089	0.0063	-1.42
Out-of-vehicle time (min)	-0.0308	0.0106	-2.90
Auto out-of-pocket cost (c)	-0.0115	0.0026	-4.39
Transit fare	-0.0070	0.0038	-1.87
Auto ownership (specific to auto mode)	-0.770	0.213	3.16
Downtown workplace (specific to auto mode)	-0.561	0.306	-1.84
Number of observations	1476		
Number of cases	1476		
LL(0)	-1023		
LL(β)	-347.4		
-2[LL(0)-LL(β)]	1371		
ρ^2	0.660		
$\bar{\rho}^2$	0.654		

Table adapted from Ben-Akiva and Lerman (1985)

← **Coefficients are not directly interpretable.** We can only interpret the effect direction, or use them to calculate utilities, and choice probabilities

To make some sense of these parameters we must calculate elasticities, marginal effects or other quantity of interest such as marginal rates of substitution (i.e. VoTT)

Basic Inference with discrete choice models

MNL: Logit Elasticities (Point elasticities)

- **Direct elasticity:** measures the **percentage change in the probability** of choosing a particular alternative in the choice set with respect to a given **percentage change** in an attribute of that same alternative.

$$E_{x_{ink}}^{P(i)} = \frac{\partial P_n(i)}{\partial x_{ink}} \cdot \frac{x_{ink}}{P_n(i)} = [1 - P_n(i)] x_{ink} \beta_k$$

- **Cross elasticity:** measures the **percentage change in the probability** of choosing a particular alternative in the choice set with respect to a given **percentage change** in a competing alternative.

$$E_{x_{jnk}}^{P(i)} = \frac{\partial P_n(i)}{\partial x_{jnk}} \cdot \frac{x_{jnk}}{P_n(i)} = -P_n(j) x_{jnk} \beta_k$$

← Because of IIA, cross-elasticities are uniform across all alternatives

Basic Inference with discrete choice models

MNL: Logit Elasticities (Point elasticities)

$when\ x_{ink} = f^k(z_{ink})$

- **Direct elasticity:** measures the **percentage change in the probability** of choosing a particular alternative in the choice set with respect to a given **percentage change** in an attribute of that same alternative.

$$E_{x_{ink}}^{P(i)} = [1 - P_n(i)]\beta_k \cdot \frac{\partial f^k}{\partial z_{ink}} z_{ink}$$

As such, when $x_{ink} = ln(z_{ink})$

$$E_{x_{ink}}^{P(i)} = [1 - P_n(i)]\beta_k \cdot \frac{\partial \ln(z_{ink})}{\partial z_{ink}} z_{ink} = \boldsymbol{[1 - P_n(i)]\beta_k}$$

Basic Inference with discrete choice models

MNL: Logit Elasticities (Point elasticities)

- The elasticities shown before are individual elasticities (Disaggregate)
- To calculate sample (aggregate) elasticities we use the **probability weighted sample enumeration** method:

$$E_{x_{ink}}^{\overline{P(i)}} = \frac{\sum_{n=1}^{N} \hat{P}_n(i) E_{x_{ink}}^{P(i)}}{\sum_{n=1}^{N} \hat{P}_n(i)}$$

Sample direct elasticity

$$E_{x_{jnk}}^{\overline{P(i)}} = \frac{\sum_{n=1}^{N} \hat{P}_n(i) E_{x_{jnk}}^{P(i)}}{\sum_{n=1}^{N} \hat{P}_n(i)}$$

Sample cross-elasticity

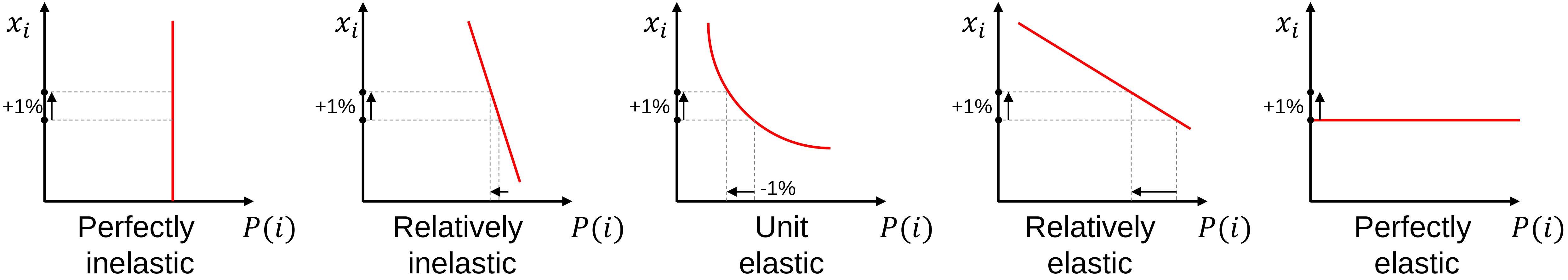
Where $\overline{P(i)}$ *is the aggregate choice probability of alternative I, and* $\hat{P}_{in}(i)$ *is an estimated choice probability*

- Uniform cross-elasticities do not necessarily hold at the aggregate level
- Also note that elasticities for dummy variables are **meaningless!**

Basic Inference with discrete choice models

Graphical illustration of elasticities

Let x_i be the cost of alternative i



Direct elasticity:

Perfectly inelastic	Relatively inelastic	Unit elastic	Relatively elastic	Perfectly elastic
1% increase in x_i results in a 0% change in $P(i)$	1% increase in x_i results in a less than 1% decrease in $P(i)$	1% increase in x_i results in a 1% decrease in $P(i)$	1% increase in x_i results in a more than 1% decrease in $P(i)$	1% increase in x_i results in a ∞ percent decrease in $P(i)$

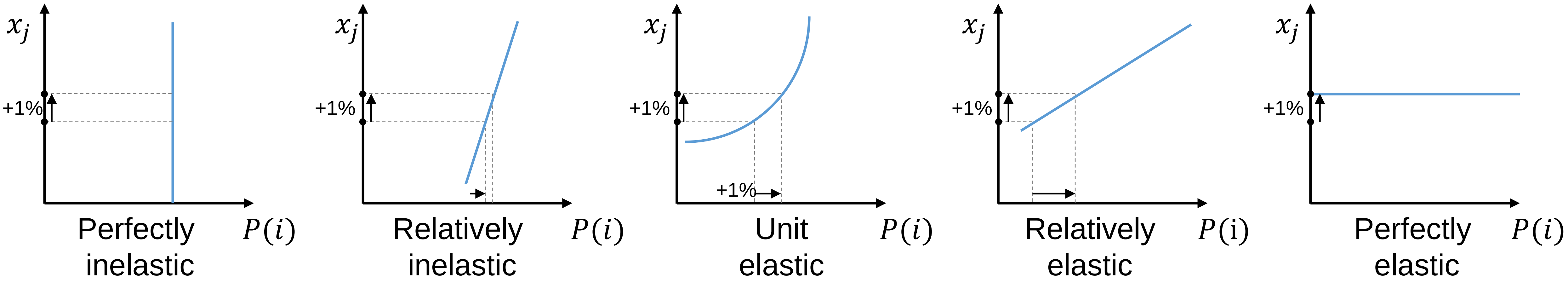
Cross elasticity:

Perfectly inelastic	Relatively inelastic	Unit elastic	Relatively elastic	Perfectly elastic
1% increase in x_j results in a 0% change in $P(i)$	1% increase in x_j results in a less than 1% increase in $P(i)$	1% increase in x_j results in no percent change in $P(i)$	1% increase in x_j results in a more than 1% increase in $P(i)$	1% increase in x_j results in a ∞ percent increase in $P(i)$

Basic Inference with discrete choice models

Graphical illustration of elasticities

Let x_j be the cost of alternative j



Direct elasticity:

1% increase in x_i results in a 0% change in $P(i)$

1% increase in x_i results in a less than 1% decrease in $P(i)$

1% increase in x_i results in a 1% decrease in $P(i)$

1% increase in x_i results in a more than 1% decrease in $P(i)$

1% increase in x_i results in a ∞ percent decrease in $P(i)$

Cross elasticity:

1% increase in x_j results in a 0% change in $P(i)$

1% increase in x_j results in a less than 1% increase in $P(i)$

1% increase in x_j results in no percent change in $P(i)$

1% increase in x_j results in a more than 1% increase in $P(i)$

1% increase in x_j results in a ∞ percent increase in $P(i)$

Basic Inference with discrete choice models

MNL: Marginal Effects

- **Direct marginal effect:** measures the **change in the probability (absolute change)** of choosing a particular alternative in the choice set with respect to a **unit change** in an attribute of that same alternative.

$$M_{x_{ink}}^{P(i)} = \frac{\partial P_n(i)}{\partial x_{ink}} = P_n(i)[1 - P_n(i)]\beta_k$$

- **Cross marginal effect:** measures the **change in the probability (absolute change)** of choosing a particular alternative in the choice set with respect to a **unit change** in a competing alternative.

$$M_{x_{jnk}}^{P(i)} = \frac{\partial P_n(i)}{\partial x_{jnk}} = P_n(i)(-P_n(j)\beta_k)$$

Basic Inference with discrete choice models

MNL: Marginal Effects

- We can also calculate sample (aggregate) marginal effects using the **probability weighted sample enumeration** method:

$$M_{x_{ink}}^{\overline{P(i)}} = \frac{\sum_{n=1}^{N} \hat{P}_n(i) M_{x_{ink}}^{P(i)}}{\sum_{n=1}^{N} \hat{P}_n(i)}$$

Sample direct marginal effect

$$M_{x_{jnk}}^{\overline{P(i)}} = \frac{\sum_{n=1}^{N} \hat{P}_n(i) M_{x_{jnk}}^{P(i)}}{\sum_{n=1}^{N} \hat{P}_n(i)}$$

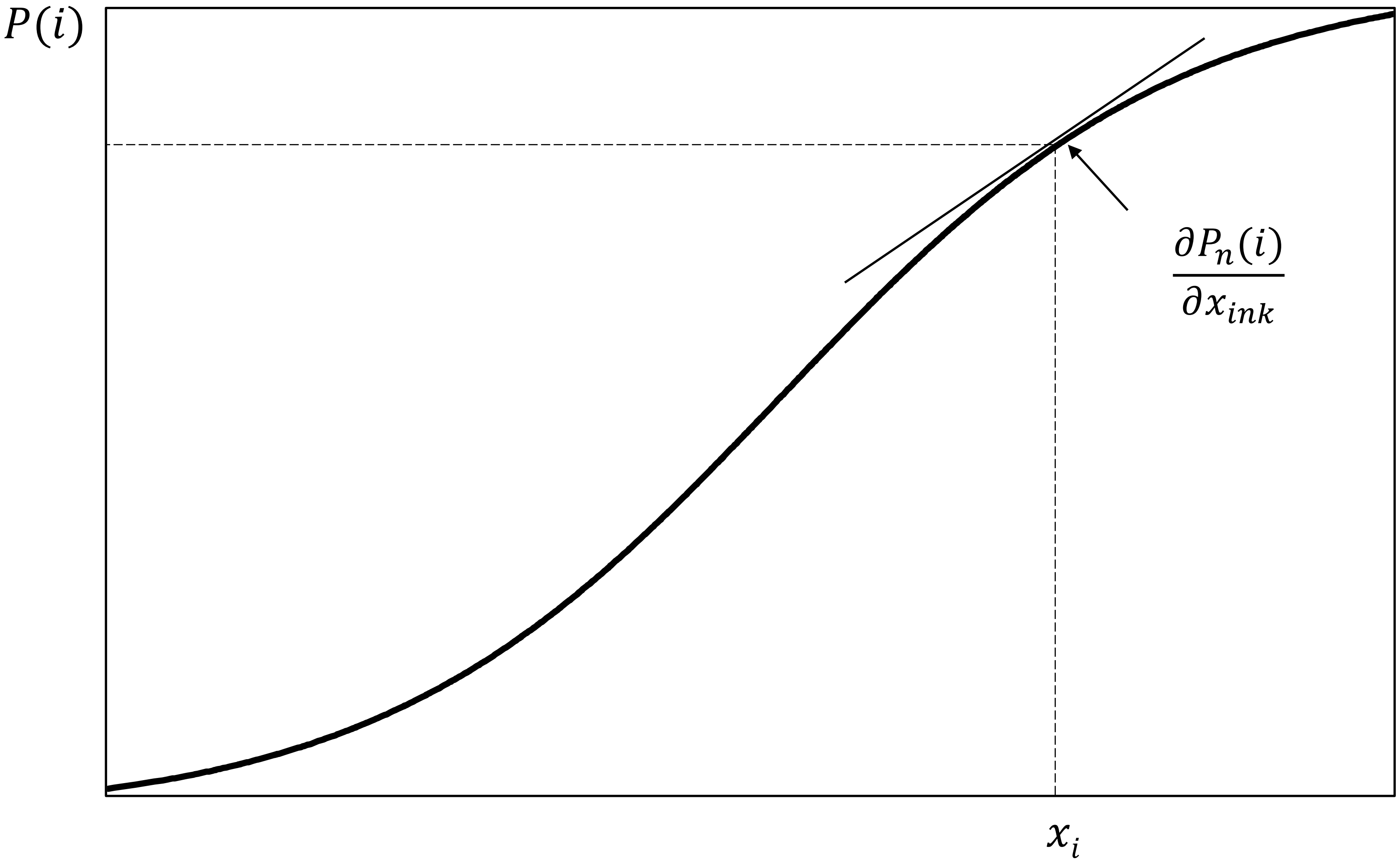
Sample cross-marginal effect

Where $\overline{P(i)}$ is the aggregate choice probability of alternative I, and $\hat{P}_{in}(i)$ is an estimated choice probability

- Marginal effects for dummy variables **do make sense** as we are talking about unit changes, but a different procedure is necessary to estimate marginal effects.

Basic Inference with discrete choice models

MNL: Marginal Effects



Marginal effects as the slopes of the Tangent lines to the cumulative probability curve

Adapted from Hensher, David A., John M. Rose, and William H. Greene. *Applied choice analysis: a primer*. Cambridge University Press, 2015 (2nd Edition)

Basic Inference with discrete choice models

MNL: Marginal Effects

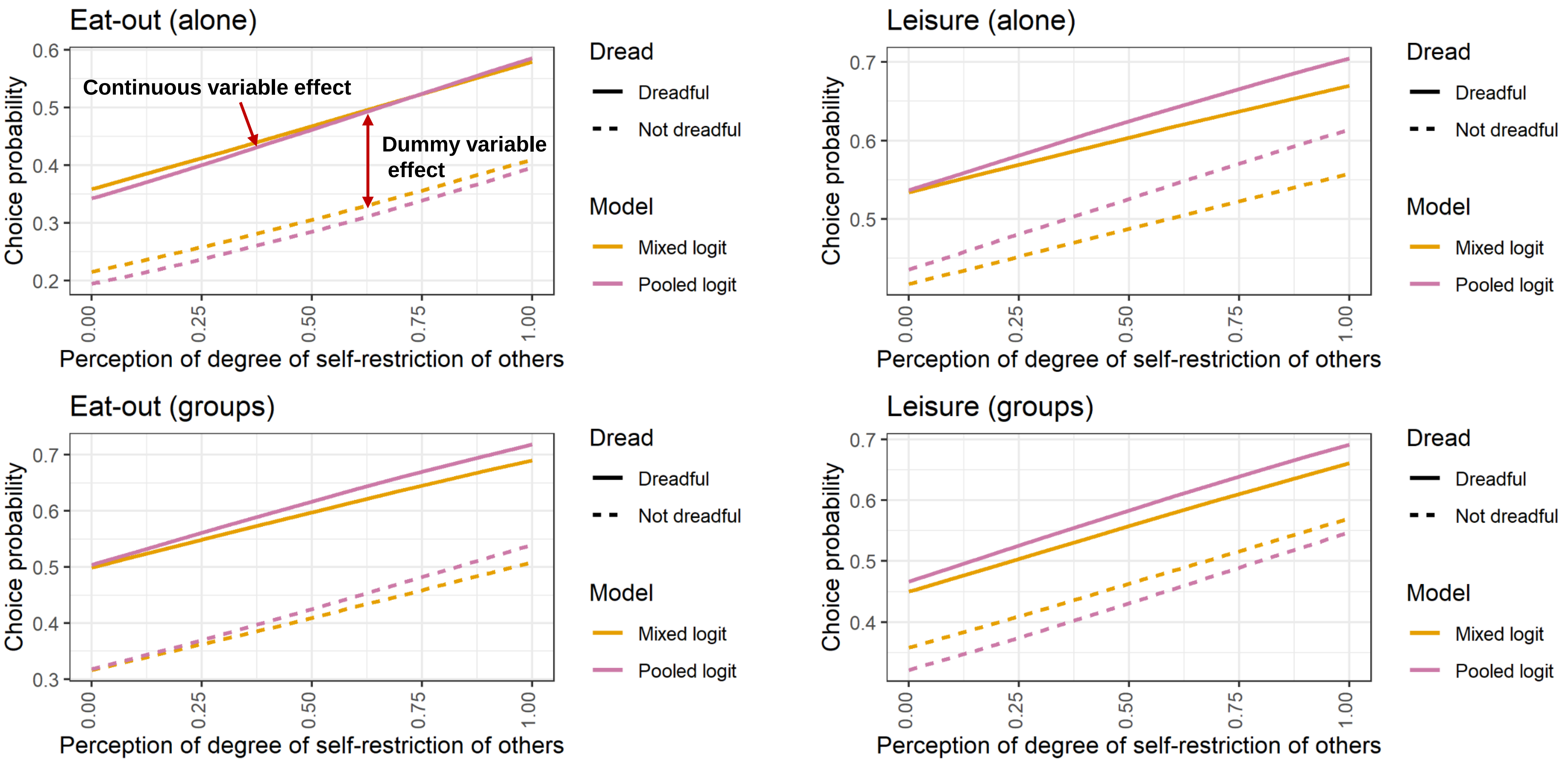
Calculating marginal effects for dummy variables

Calculated via simulation:

1. Set the values of the variable of interest to 0
2. Estimate base predictions (at the individual level)
3. Set the values of the variable of interest to 1
4. Estimate new predictions (at the individual level)
5. Calculate marginal effects by taking the mean of the difference in individual predictions

Basic Inference with discrete choice models

Simulation and visualization of estimation results



Simulation of the effects of perception of degree of self-restriction of others, and COVID-19 dread on going-out self-restriction (“stay home”) choice probability for eating-out and leisure, and comparison between binary logit and mixed logit results

Other covariates are fixed as follows: time period = t_1. All continuous variable set to mean values. All categorical variables set to reference categories

Rather obvious…

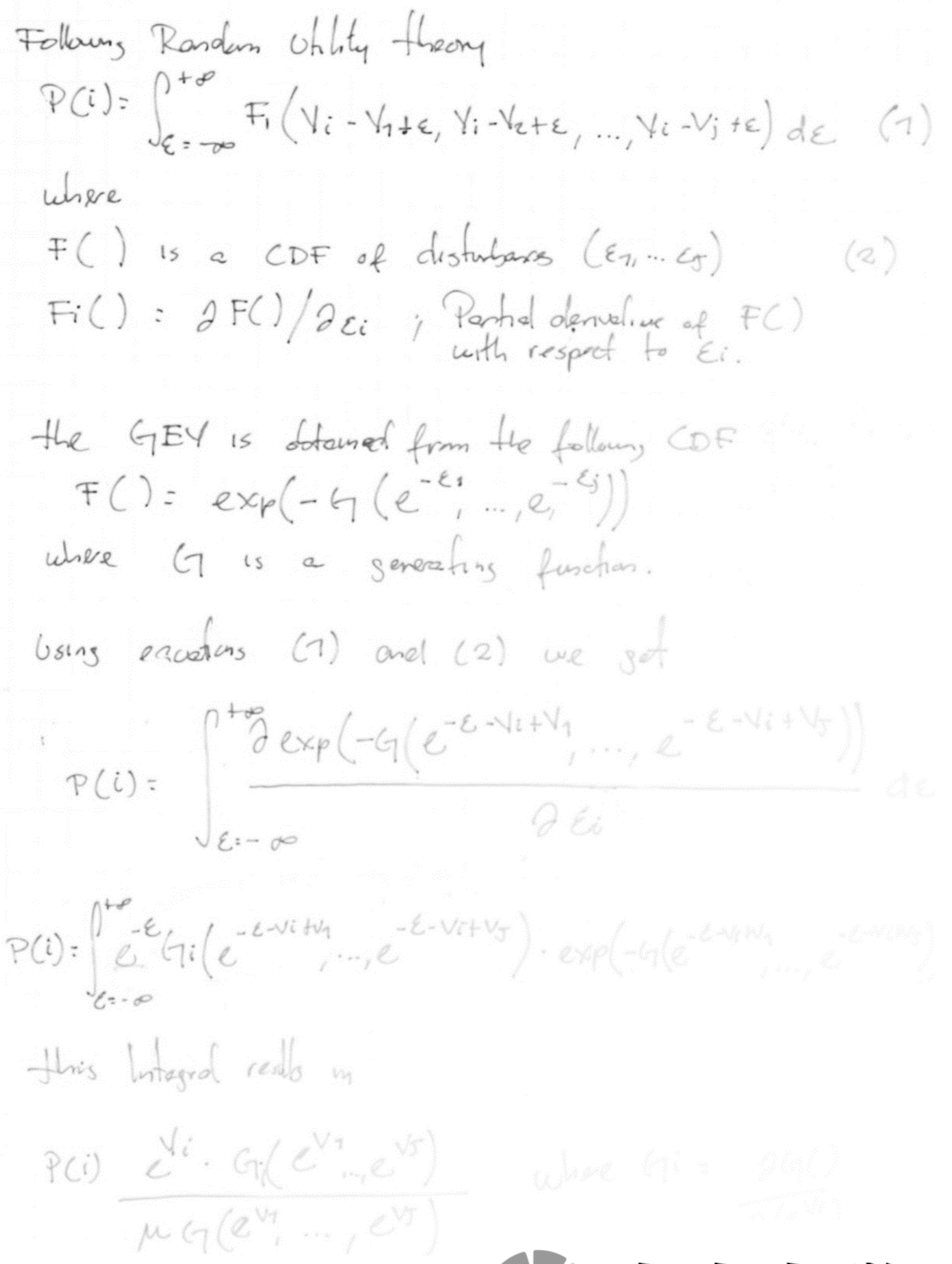
Is it though?

Size Matters:

The Use and Misuse of Statistical Significance in Discrete Choice Models

The problem: Quantitative researchers are obsessed with statistical significance.

Where does the field stand regarding the use and misuse of statistical significance in empirical analysis?



東京大学
THE UNIVERSITY OF TOKYO

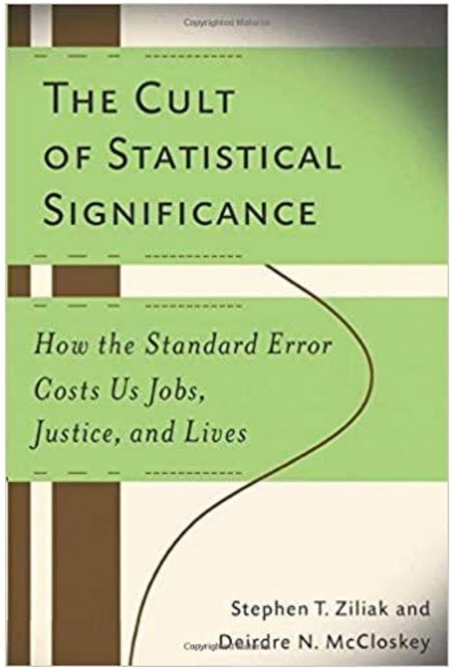
The Use and Misuse of Statistical Significance in Discrete Choice Models

A review of reporting practices in the Transportation Academic Literature

- Follow-up study to the work of Parady et al., (2021) that showed that while 92% of studies reported goodness-of-fit statistics, only 18.1% reported validation.

- Based on the seminal work of McCloskey and Ziliak (1996) in economics.

- We adapt McCloskey and Ziliak (1996)'s 19 questions to the academic transportation literature to evaluate where the field stands regarding the use and misuse of statistical significance in empirical analyses



Does the article... (Think about the last article you wrote)	% Yes	Out of which:	
		Comprehensively	Limitedly
Q4: **Consider the power** of the test?	0.00	-	-
Q5: Examine the power function?	-	-	-
Q15: Report **effect confidence intervals**, using them to interpret economic significance not merely as a replacement for pointwise statistical significance?	7.37	0	7.37
Q10: Discuss the **scientific conversation** within which an effect or other quantity of interest would be judged large or small?	13.68	-	-
Q12**: Do a simulation** to determine whether the estimated effects or other quantities of interest are reasonable and/or to better illustrate the magnitude of estimated effects?	29.47	-	-
Q13: In the conclusions and implications sections, **keep statistical significance separate from economic policy and scientific significance?**	32.63	-	-
Q9: Make a **judgement on effect magnitudes**?	36.84	13.68	23.16
Q14: In the estimation, conclusions, and implication sections**, avoid using the word "significance" in ambiguous ways**?	37.63	-	-
Q7: In the model results section, **eschew "sign econometrics**"?	60.64	27.66	32.98
Q8**: Discuss the magnitude of estimated effects** or other quantities of interest?	64.21	33.68	30.53
Q2: Use coefficients **to calculate elasticities, or some other quantity** that addresses the question of "how large is large"?	65.26	45.26	20.00
Q11**: Avoid choosing variables for inclusion solely on the basis of statistical significance**?	75.53	-	-
Q3: Report all traditionally reported statistics?	76.84	-	-
Q1: Report descriptive statistics for model variables?	78.95	65.26	13.68
Q6: Eschew "asterisk econometrics"?	100.00	-	-

The Use and Misuse of Statistical Significance in Discrete Choice Models

A review of reporting practices in the Transportation Academic Literature

67% of reviewed studies **did not distinguish statistical significance from economic, policy or scientific significance.**

86% of studies **did not discuss the scientific conversation within which the magnitude of a coefficient can be judged to be "large" or "small."**

62% of studies **ambiguously used the word "significant"** to mean statistically different from the null sometimes and to mean practically important at other times.

39% explained model results exclusively based on the sign of the coefficient.

24% explicitly stated to have **used statistical significance as an exclusive criterion to drop variables** from a model.

0% of the reviewed studies considered the statistical power of the tests.

The Use and Misuse of Statistical Significance in Discrete Choice Models

A review of reporting practices in the Transportation Academic Literature

DO NOT! **Discuss your findings only in terms of significance and sign (sign econometrics)**

Statistical significance **is just a measure of ONE kind of error**, but it does not tell you anything about whether that effect is practically important

Plus or minus, who cares? **Without a measure of size, this information is useless!**

"Variable X has a significant and positive effect on…"

Statistical significance asks whether an effect exists, not how big is it.

With a large enough sample, EVERYTHING is significant

The Use and Misuse of Statistical Significance in Discrete Choice Models

A review of reporting practices in the Transportation Academic Literature

Good Practices: **Discussing effect magnitude (not significance!) and making judgement about how large an effect is**

Khan, Kockelman and Xiong (*2014*) make clear judgements of magnitude when they state that "*network connectivity (measured as 4-way intersections within 0.5 mile) plays a major role: a single standard deviation change in this variable is estimated to increase walking probability by 34%*" and go on to state that "*parking prices and free-parking availability variables were not found to have much of an effect.*"

de Luca and Di Pace (2015) also make clear judgments of magnitude when they discuss the magnitude of value of travel time estimates and state that *"aside from being similar to those estimated in different Italian case studies, [the magnitude] indicates the extreme importance of parking location. Assuming that the average one-way travel monetary cost is equal to 3 €, 10 min walking time (about 700 m at 4 km/h) is more than half of the whole travel monetary cost."*

The Use and Misuse of Statistical Significance in Discrete Choice Models

A review of reporting practices in the Transportation Academic Literature

Other common mistakes: **Confusing statistical significance with practical importance**

Kamargianni et al. (2014) state of a latent construct of walking preference that *“this component is the most statistically significant variable…indicating the strong influence that parents have on the development of their children’s attitudes towards walking”*

Qin et al. (2017) argue in a study of mode-shifting behavior that *“bus service level has the most significant positive t-value, which indicates that improving the bus service level can increase the shifting proportion of car travelers to bus significantly.”*

The Use and Misuse of Statistical Significance in Discrete Choice Models

A review of reporting practices in the Transportation Academic Literature

Recommendations:

Always report of effect magnitudes and their confidence intervals (it should be mandatory).

Statistical significance should not be more than one of many criteria of evaluation, but it should certainly not be the most important one. The discussion of statistical models should focus on effect magnitude and other policy relevant quantities. **Is it large enough to matter for policy?**

Provide to the extent possible judgements of magnitude that convey what the authors consider are "small," "medium," or "large" effects (or other quantities of interest) and the basis for such judgement.

This is certainly not an easy task, there is a discussion to be had regarding what effects or quantities are policy relevant and how to assess such relevance.

Furthermore, **such discussions should ideally be accompanied by a discussion on the cost implications** of changing the policy variables in question.

The Use and Misuse of Statistical Significance in Discrete Choice Models

A review of reporting practices in the Transportation Academic Literature

Recommendations:

Compare, whenever possible, effect magnitudes or other quantities of interest to existing studies.

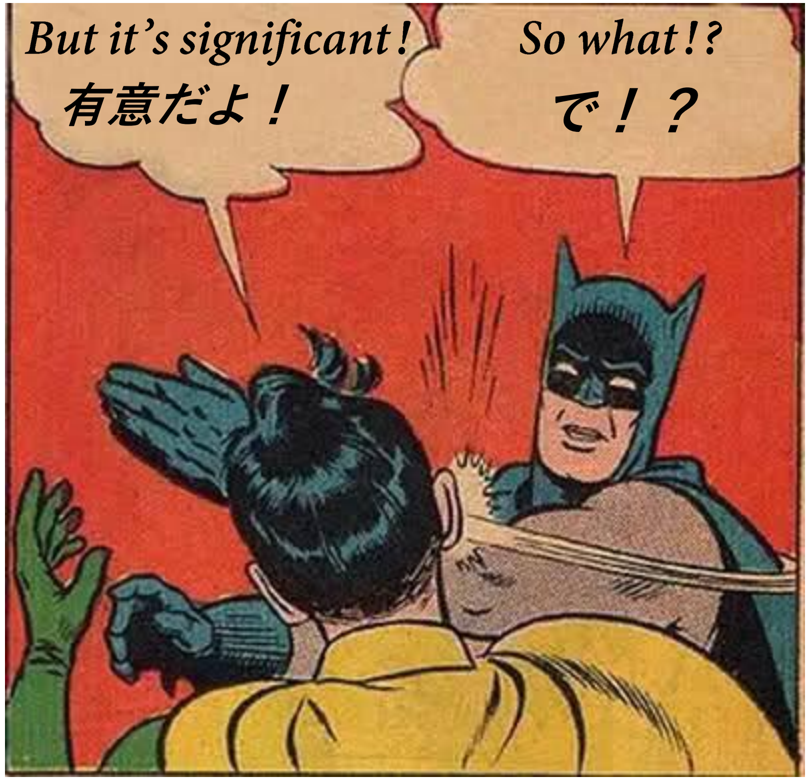
For the most regularly reported values, such as value of travel time, there is a myriad of studies reporting such values for many contexts, so there are no reasons why such comparisons cannot be made. For less often reported values, there will be certainly times when such a task will be difficult, but **if we all do it, in time, proper discussion of scientific context should be widespread.**

For new studies, take statistical power into consideration when defining sample size to guarantee the effects the researcher wants to detect can in fact be detected with enough power.

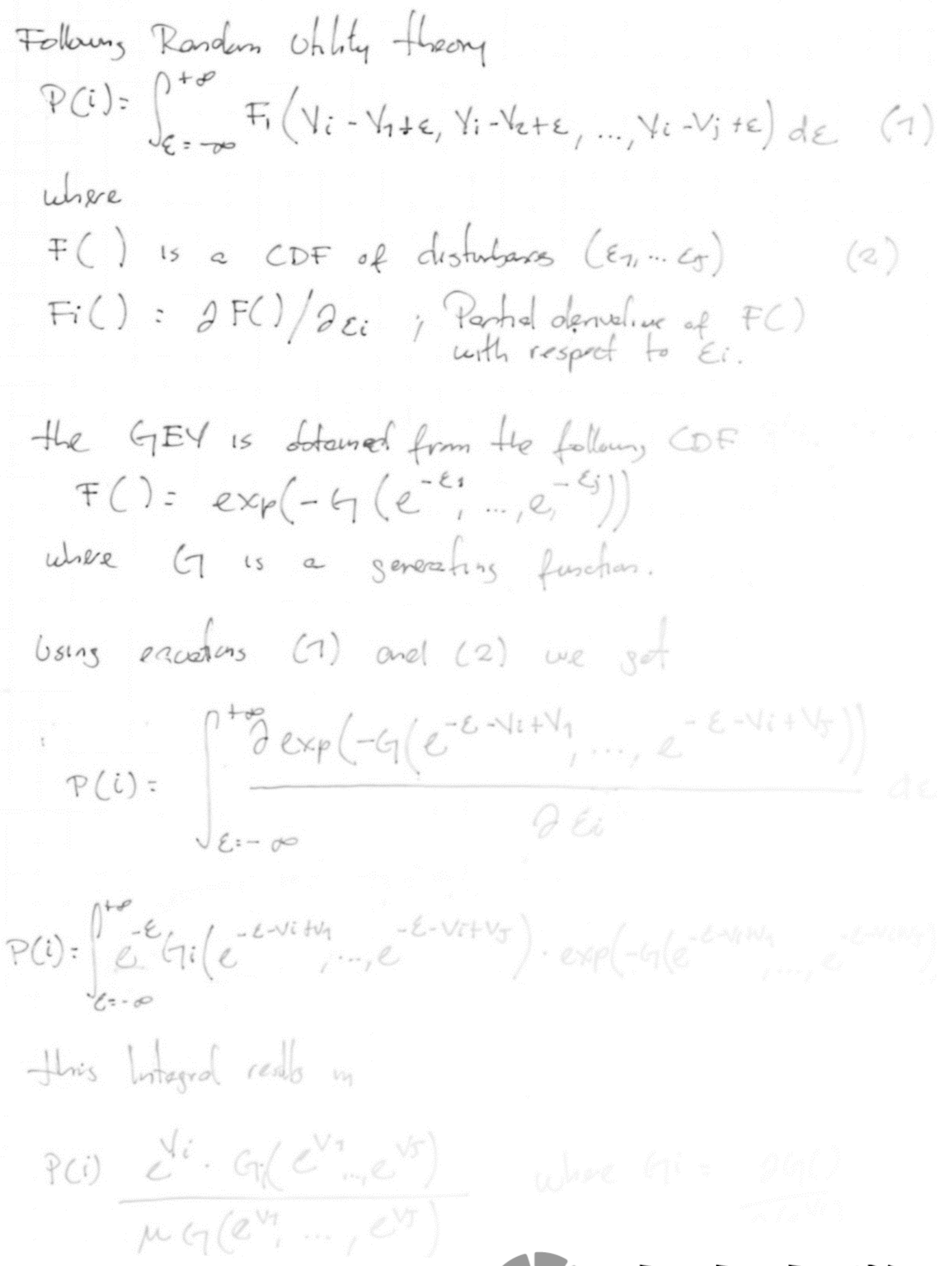
For studies using secondary data (i.e., national household survey data, etc.) report post-hoc power levels of tests reported in the study.

The 22nd Behavior Modeling Summer School
Sep. 18 – 20 , 2023 @ The University of Tokyo

Size Matters:
The Use and Misuse of Statistical Significance in Discrete Choice Models



To summarize…

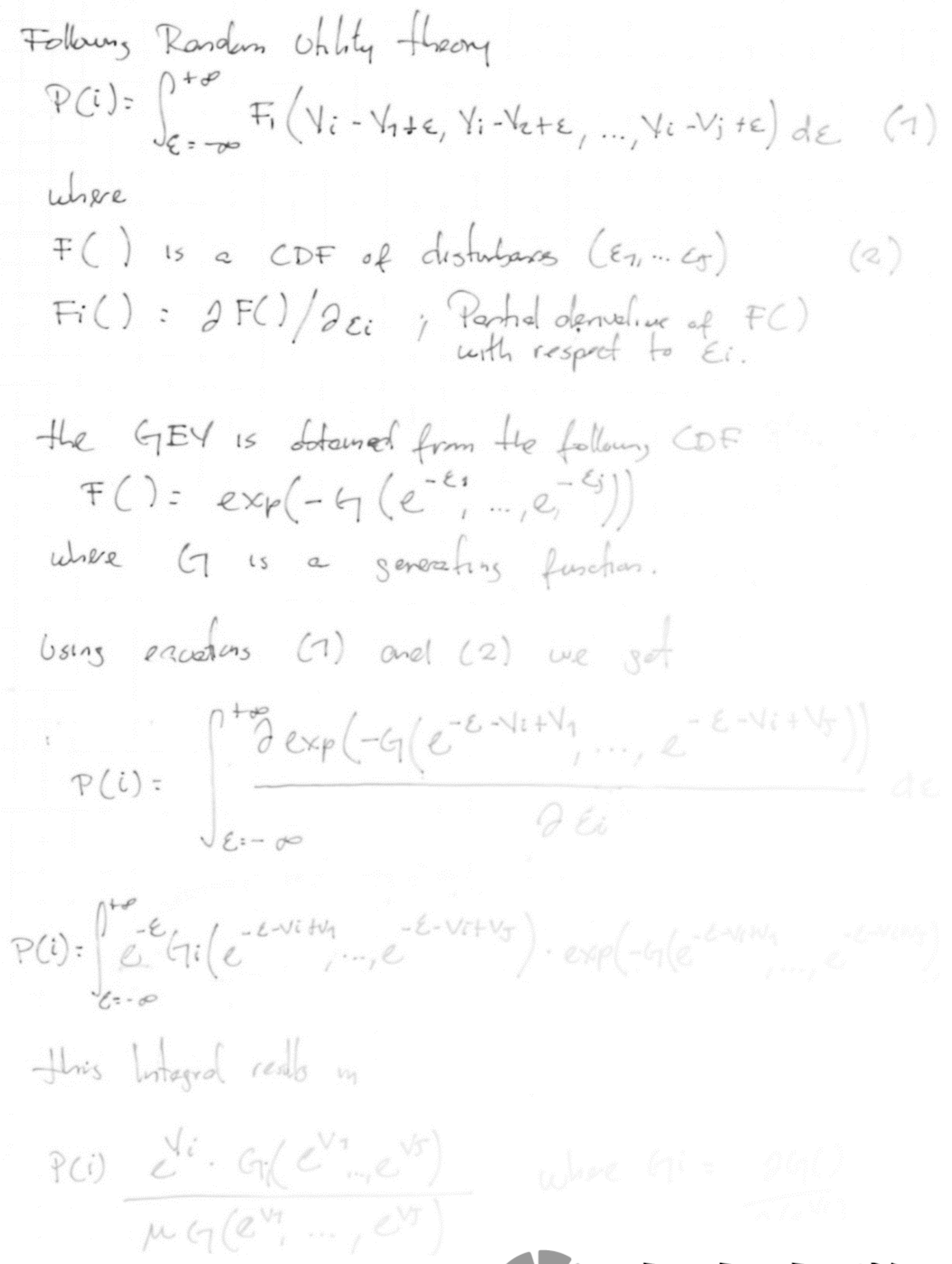


Read the full paper here: **Parady, G.**, Axhausen K.W (2023) .: Size Matters The Use and Misuse of Statistical Significance in Discrete Choice Models in the Transportation Academic Literature. Transportation (accepted)

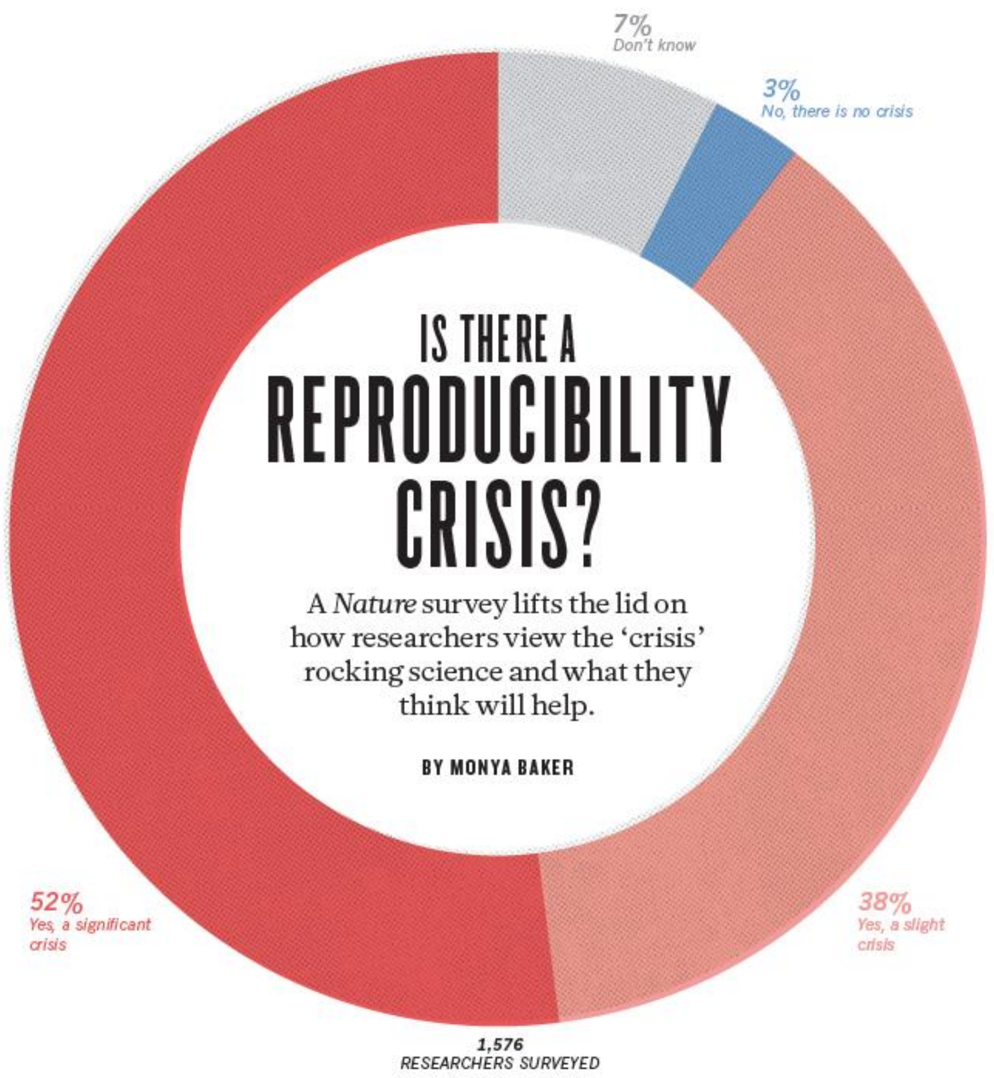
東京大学
THE UNIVERSITY OF TOKYO

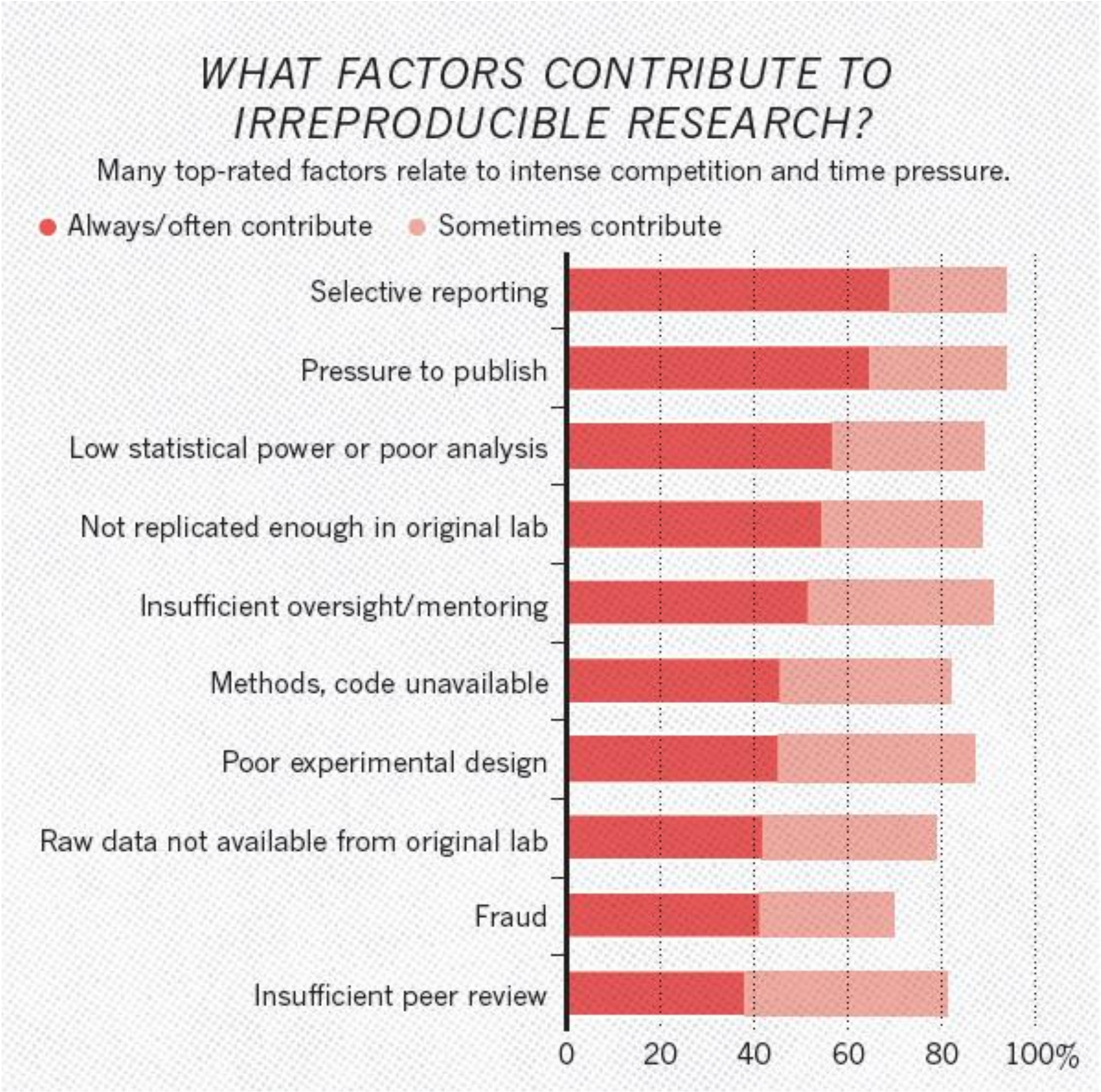
For self-study:

Validation practices in discrete choice modeling



A credibility crisis in science and engineering?





Baker and Penny (2016)

A credibility crisis in science and engineering?

Most published research findings are likely to be false due to factors such as lack of power of the study, small effect sizes, and great flexibility in research design, definitions, outcomes and methods.

Focused on experimental studies

(Ioannidis, 2005)

In the transportation field

Unlike the natural sciences

- Dependence on cross-section observational studies
- Classic scientific hypothesis testing is more difficult
- Impact evaluation of policies drawn based on model-based academic research is rarely conducted
- No feedback in terms of how right or how wrong are these models and the policy recommendations derived from them
- **These issues underscore the need for proper validation practices**

Term definitions

Predictive accuracy: The degree to which predicted outcomes match observed outcomes.

Predictive accuracy is a function of：

- **Calibration:** The degree to which predicted probabilities match the relative frequency of observed outcomes.
- **Discrimination ability:** The ability of a model or system of models to discriminate between those instances with and without a particular outcome.

Generalizability: The ability of a model, or system of models to maintain its predictive accuracy in a different sample.

Generalizability of a model is a function of：

- **Reproducibility:** The extent to which a model or system of models maintains its predictive ability in different samples from the same population.
- **Transferability:** The extent to which a model or system of models maintains its predictive ability in samples from different but plausibly related populations or in samples collected with different methodologies (sometimes called transportability)

Parady, Ory & Walker (2021)

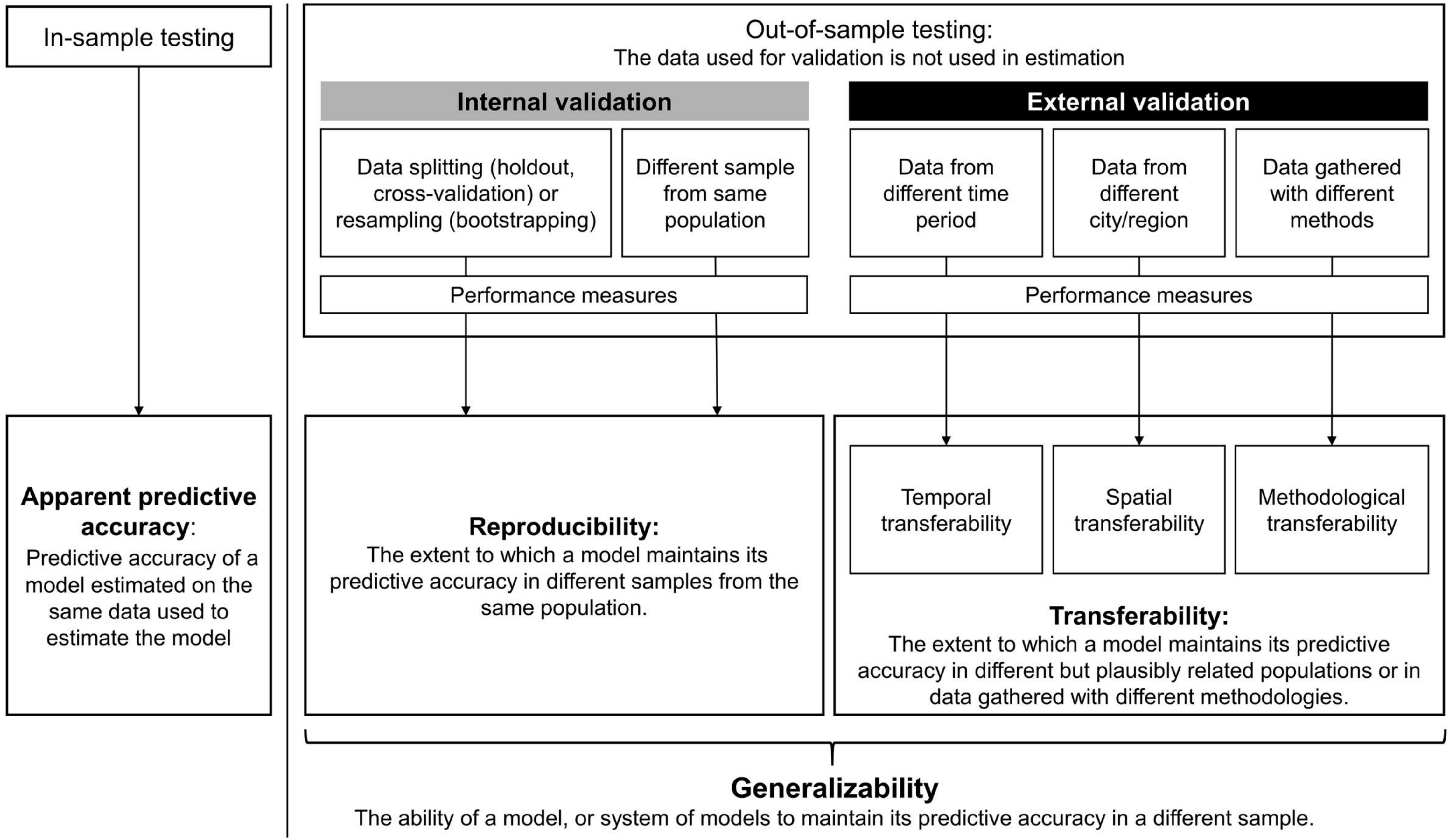
Term definitions

Model validation: The evaluation of the generalizability of a statistical model.

Types of model validation：

- **Internal validation:** The evaluation of the reproducibility of a model.
  - Data splitting (i.e., cross-validation), resampling methods (i.e., bootstrapping)
  - Different sample from the same population
- **External validation:** The evaluation of the transferability of a model.
  - Temporal transferability
  - Spatial transferability
  - Methodological transferability

Parady, Ory & Walker (2021)



Parady, Ory & Walker (2021)

A brief introduction to internal validation (data splitting methods)

Holdout validation: Dataset is randomly split into an estimation dataset and a validation dataset.

For illustration purposes, let us define $\mathrm{Q}[y_n, \hat{y}_\mathrm{n}]$ as a measure of prediction correctness for the nth instance, for the binary choice case as:

$$\mathrm{Q}[y_n, \hat{y}_\mathrm{n}] = \begin{cases} 0 \; if \; y_n = \hat{y}_\mathrm{n} \\ 1 \; if \; y_n \neq \hat{y}_\mathrm{n} \end{cases}$$

where is y_n the observed outcome, and $\hat{y}_\mathrm{n}$ is the predicted outcome for instance *n*.

The holdout estimator is

$$HOV = \frac{1}{N_v} \sum_{n_v=1}^{N_v} Q[y_{n_v}, \hat{y}^e_{n_v}]$$

where $\hat{y}^e_{n_v}$ is the predicted outcome for instance n in sample v, using the model estimated with sample e, and N_v is the validation sample size.

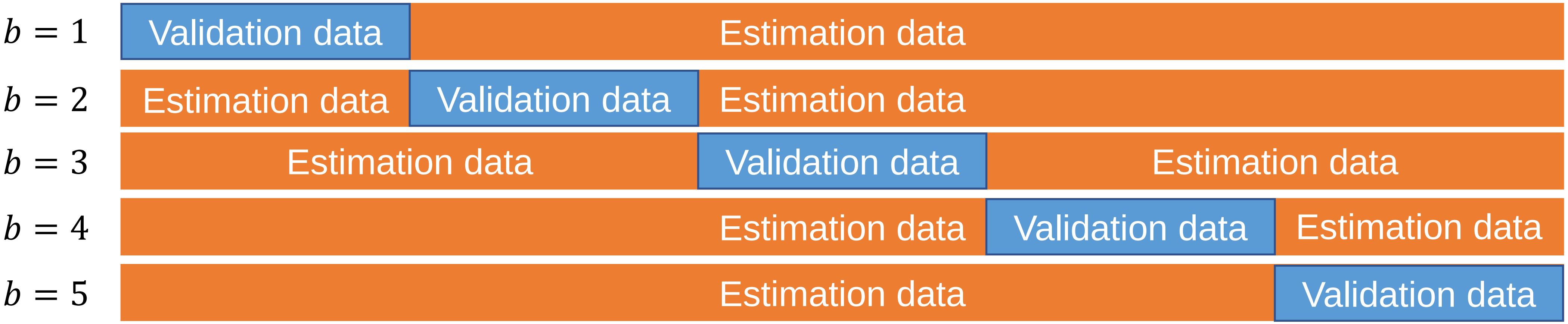
Parady, Ory & Walker (2021)

A brief introduction to internal validation (data splitting methods)

Cross-validation: When the holdout process is repeated multiple times, thus generating a set of randomly split estimation-validation data pairs, we refer to the validation procedure as cross-validation (CV).

$$CV = \frac{1}{\mathrm{B}}\sum_{b} HOV_b$$

where B is the number of estimation-validation data pairs generated and is the holdout estimator for set *b.*



A 5-fold cross validation illustration

Parady, Ory & Walker (2021)

A brief introduction to internal validation (data splitting methods)

Cross-validation：Commonly used methods

$$CV = \frac{1}{B}\sum_{b} HOV_b$$

- Cross-validation **methods differ from one another in the way the data is split.**
- When the data splitting considers all possible estimation sets of size , the splitting is **exhaustive**, otherwise the splitting is **partial**. (Arlot and Celisse, 2009)。

Exhaustive splitting methods

- **Leave-one-out :** estimation set size is $\mathrm{N_e} = N - 1$, and B = N. The model is fitted leaving out one instance per iteration, and the outcome of that single instance is predicted based on the estimated model.
- **Leave-p-out :** $\mathrm{N_e} = N - p$. The model is fitted leaving out p-instances per iteration, and the outcome of the remaining instances is predicted based on the estimated model.

Parady, Ory & Walker (2021)

A brief introduction to internal validation (data splitting methods)

Cross-validation：Commonly used methods

$$CV = \frac{1}{B}\sum_{b} HOV_b$$

- Cross-validation **methods differ from one another in the way the data is split.**
- When the data splitting considers all possible estimation sets of size , the splitting is **exhaustive**, otherwise the splitting is **partial**. (Arlot and Celisse, 2009)。

Partial splitting methods (lower calculation cost)

- **K-fold cross-validation:** data is partitioned into *K* mutually-exclusive subsets of roughly equal size, and *B*=*K.*
- **Repeated learning-testing:** a B number of randomly-split estimation-validation pairs are generated. This method is also called repeated holdout validation.

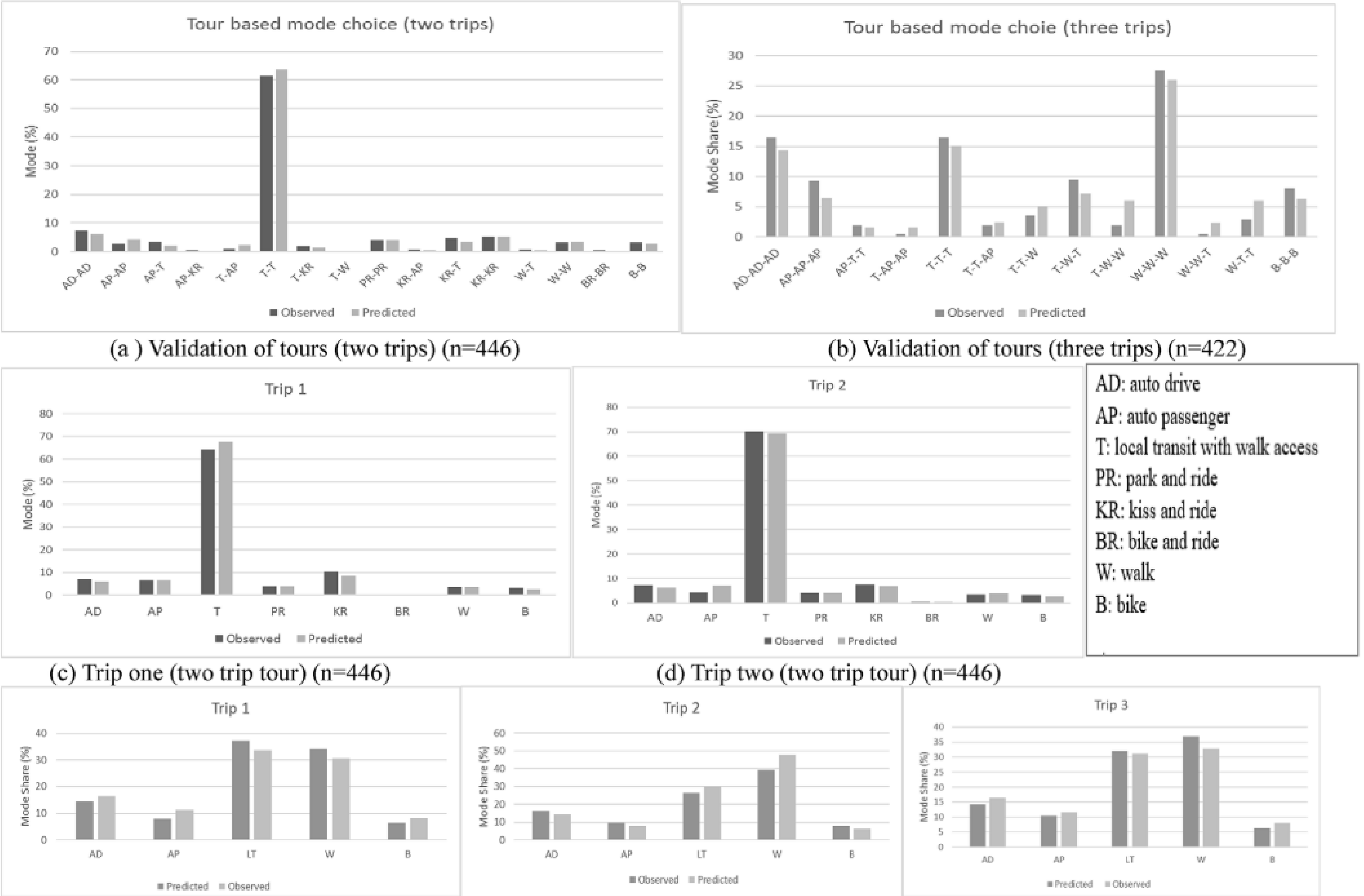
Parady, Ory & Walker (2021)

A brief introduction to internal validation (data splitting methods)

Performance measures

Market share comparison

- Easy to execute
- Does not provide a quantitative measure to evaluate the level of agreement between predictions and observations



(a) Validation of tours (two trips) (n=446)
(b) Validation of tours (three trips) (n=422)
(c) Trip one (two trip tour) (n=446)
(d) Trip two (two trip tour) (n=446)
(e) Trip one (three trip tour) (n=422)
(f) Trip two (three trip tour) (n=422)
(f) Trip three (three trip tour) (n=422)

Fig. 3. Validation results of trips and tours.

Hasnine and Habib (2018)

A brief introduction to internal validation (data splitting methods)

Performance measures

Percentage of correct predictions: the alternative with the highest probability is defined as the predicted choice. However,

Model A :

- **Alt. A: 0.34＊**
- Alt. B: 0.33
- Alt. C: 0.33

Model B :

- **Alt. A: 0.50＊**
- Alt. B: 0.30
- Alt. C: 0.20

Model C :

- **Alt. A: 0.90＊**
- Alt. B: 0.05
- Alt. C: 0.05

＊Observed choice

Cannot discriminate differences in estimated probabilities.

A measures that accounts for “clearness” of prediction is necessary.

A brief introduction to internal validation (data splitting methods)

Performance measures

Clearness of prediction:

Percentage of clearly right choices: *"the percentage of users in the sample whose observed choices are given a probability greater than threshold t by the model"*

$$\%\mathrm{CR} = \frac{100}{N_v}\sum_{n_v=1}^{N_v} CR_{n_v} \quad where \quad \mathrm{CR}_{n_v} = \begin{cases} 1\ if\ \hat{P}\left(y_{n_v}^{e}\right) > t \\ 0 \quad otherwise \end{cases}$$

Percentage of clearly wrong choices: *"the percentage of users in the sample for whom the model gives a probability greater than threshold t to a choice alternative differing form the observed one"*

$$\%\mathrm{CW} = \frac{100}{N_v}\sum_{n_v=1}^{N_v} CW_{n_v} \quad where \quad CW_{n_v} = \begin{cases} 1\ if\ \hat{P}\left(!\,y_{n_v}^{e}\right) > t \\ 0 \quad otherwise \end{cases}$$

$\hat{P}\left(!\,y_{n_v}^{e}\right)$ is the estimated choice probability of an alternative other than the chosen one.

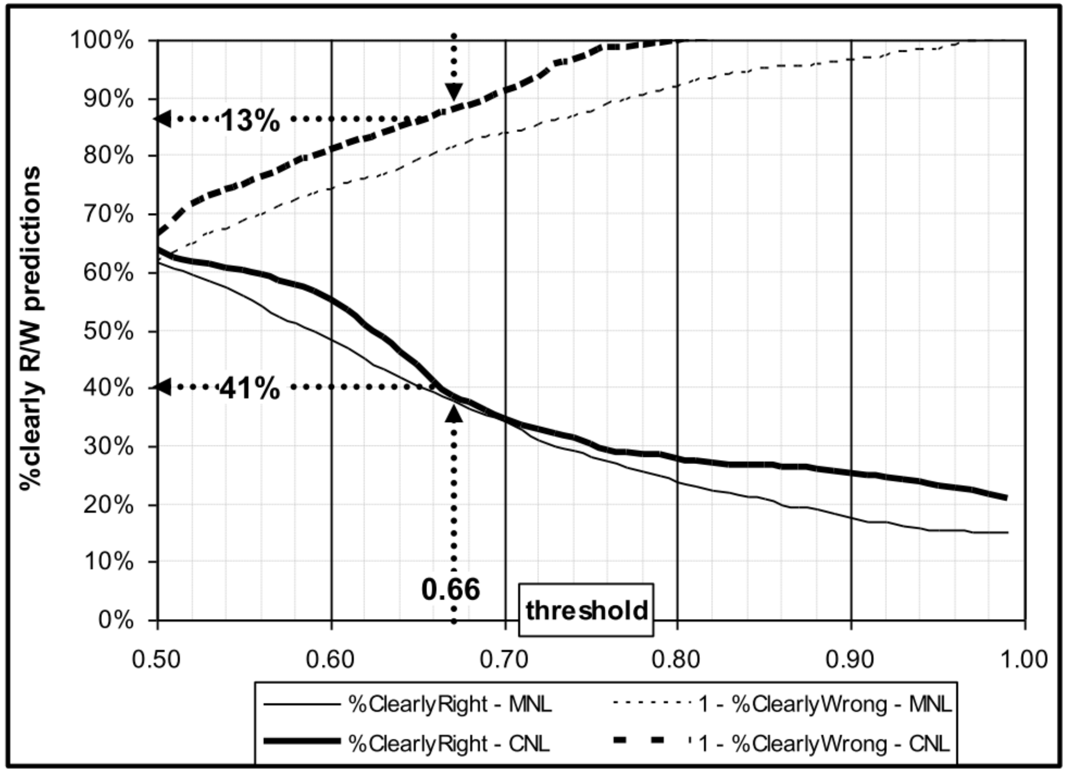
de Luca and Cantarella (2009)

A brief introduction to internal validation (data splitting methods)

Performance measures

Clearness of prediction: defining threshold t

- To be meaningful, the threshold t must be "considerably larger" than c^{-1}, where c is the choice set size.
- Values used in the literature:
  - Binary model： $t = 0.9$ (de Luca and Di Pace, 2015)
  - Trinary model： $t = 0.5$ (Glerum, Atasoy and Bierlaire , 2014)



de Luca and Di Pace (2015)

See appendix for a list of commonly used indicators

Validation and reporting practices in the transportation academic literature

226 articles reviewed by Parady, Ory and Walker (2021)

92% reported a goodness of fit statistics

64.6% reported a policy-related inference

Marginal effects, elasticities, odds ratios, value of time estimates, marginal rates of substitution, and policy scenario simulations

18.1% reported a validation measure

Table 3

Internal validation methods reported in the literature by frequency.

Method	Abbvr.	Frequency	Percentage
Holdout validation	HOV	18	56.3%
Repeated learning-testing	RLT	8	25.0%
Validation against an independent sample	IS	4	12.5%
Repeated K-fold cross-validation	R-K-CV	1	3.1%
Other sample splitting methods	SS-O	1	3.1%

Towards better validation practices in the field

- **Make model validation mandatory:**
  - Non-negotiable part of model reporting and peer-review in academic journals for any study that provides policy recommendations.
  - Cross-validation is the norm in machine learning studies.

- **Share benchmark datasets:**
  - A fundamental limitation in the field is the lack of benchmark datasets and a general culture of sharing code and data.

- **Incentivize validation studies:**
  - Lot of emphasis on theoretically innovative models.
  - Encourage submissions that focus on proper validation of existing models and theories.

- **Draw and enforce clear reporting guidelines:**
  - In addition to detailed information of survey characteristics such as sampling method, discussion on representativeness of the data, validation reporting is required.
  - Efforts to improve reporting are well documented in other fields
(i.e. STROBE statement (von Elm et al., 2007))

Wait a minute…

Q: "I'm not validating my model because I'm not trying to build a predictive framework. I'm trying to learn about travel behavior"

A: **The more orthodox the type of analysis, the stronger the onus of validation.**

Q: "*Does every study that uses a discrete choice model should be conducting validation?"*

A: In short, yes. At the very least, **any article that makes policy recommendations should be subject to proper validation** given the dependence of the field on cross-section observational studies, and the lack of a feedback loop in academia.

Q: "Is what we learn about travel behavior from coefficient estimation less valuable if not conducted?"

A: There is a myriad of reasons why some **skepticism is warranted** against any particular model outcome. the most obvious one being model overfitting.

Finally

Better validation practices will not solve the credibility crisis in the field, but it’s a step in the right direction.

Model validation is **no solution to the causality problem** in the field, but we want to underscore that **the reliance on observational studies inherent to the field demands more stringent controls to improve external validity of results**.

References:

1. Ben-Akiva, M. E., Lerman, S. R. (1985). Discrete choice analysis: theory and application to travel demand. MIT press.

2. M. Baker, D. Penny (2016) Is there a reproducibility crisis? Nature, 533 (7604) pp. 452-454

3. de Luca, S. De and Cantarella, G. E. (2009) 'Validation and comparison of choice models', in Saleh, W. and Sammer, G. (eds) Travel Demand Management and Road User Pricing: Success, Failure and Feasibility. Ashgate publications, pp. 37–58. doi: 10.1017/cbo9780511619960.008.

4. de Luca, S., and R. Di Pace (2015). Modelling Users' Behaviour in Inter-Urban Carsharing Program: A Stated Preference Approach. Transportation Research Part A: Policy and Practice, Vol. 71, pp. 59–76. https://doi.org/10.1016/j.tra.2014.11.001.

5. Glerum, A., Atasoy, B. and Bierlaire, M. (2014) 'Using semi-open questions to integrate perceptions in choice models', Journal of Choice Modelling. Elsevier, 10(1), pp. 11–33. doi: 10.1016/j.jocm.2013.12.001.

6. Hasnine, M. S. and Habib, K. N. (2018) 'What about the dynamics in daily travel mode choices? A dynamic discrete choice approach for tour-based mode choice modelling', Transport Policy. Elsevier Ltd, 71(August), pp. 70–80. doi: 10.1016/j.tranpol.2018.07.011.

7. Hensher, D. A., Rose, J. M., & Greene, W. H. (2015). Applied choice analysis: a primer. Cambridge University Press. 2nd Edition.

8. Kamargianni, M., M. Ben-Akiva, and A. Polydoropoulou (2014) Incorporating Social Interaction into Hybrid Choice Models. Transportation, Vol. 41, No. 6, pp. 1263–1285.

9. Qin, H., J. Gao, H. Guan, and H. Chi. (2017) Estimating Heterogeneity of Car Travelers on Mode Shifting Behavior Based on Discrete Choice Models. Transportation Planning and Technology, Vol. 40, No. 8, pp. 914–927. https://doi.org/10.1080/03081060.2017.1355886.

10. Khan, M., K. M. Kockelman, and X. Xiong. (2014) Models for Anticipating Non-Motorized Travel Choices, and the Role of the Built Environment. Transport Policy, Vol. 35, pp. 117–126. https://doi.org/10.1016/j.tranpol.2014.05.008.

11. McCloskey, D. N., and S. T. Ziliak.(1996) The Standard Error of Regressions. Journal of Economic Literature, Vol. 34, No. 1, pp. 97–114.

12. Parady, G., Ory, D., Walker, J. (2021) The overreliance on statistical goodness of fit and under-reliance on validation in discrete choice models: A review of validation practices in the transportation academic literature . Journal of Choice Modelling 38, 100257 (Open Access)

13. Parady, G., Axhausen K.W.: Size Matters (2023) The Use and Misuse of Statistical Significance in Discrete Choice Models in the Transportation Academic Literature. Transportation (accepted)

Appendix: Definition of model validation performance measures reported in the literature

Parady, Ory & Walker (2021)

Index		Type	Formula
Mean absolute percentage error 平均絶対誤差率	MAPE	*Absolute*	$\frac{100}{M}\sum_{m=1}^{M}\left\lvert\frac{\hat{s}^{e}{}_{v,m}-s_{v,m}}{s_{v,m}}\right\rvert$
Root sum of square error 二乗平方根誤差和	RSSE	Relative	$\sqrt{\sum_{m=1}^{M}\left(\hat{s}^{e}{}_{v,m}-s_{v,m}\right)^{2}}$
Mean absolute error 平均絶対誤差	MAE	Aggregate: Relative Disaggregate: Absolute	$\frac{1}{M}\sum_{m=1}^{M}\left\lvert\hat{s}^{e}{}_{v,m}-s_{v,m}\right\rvert$
Mean squared error 平均二乗誤差	MSE	Aggregate: Relative Disaggregate: Absolute	$\frac{1}{M}\sum_{m=1}^{M}\left(\hat{s}^{e}{}_{v,m}-s_{v,m}\right)^{2}$
Root mean square error 二乗平均平方根誤差	RMSE	Aggregate: Relative Disaggregate: Absolute	$\sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{s}^{e}{}_{v,m}-s_{mv,}\right)^{2}}$
Brier Score ブライアスコア	BS	Absolute	$\frac{1}{N_v}\sum_{n_v=1}^{N_v}\sum_{m=1}^{M}\left(\hat{P}\left(y^{e}_{n_v,m}\right)-y_{n_v,m}\right)^{2}$

Notes

M is the number of alternatives in the choice set.

$s_{v,m}$ is an aggregate outcome measure in sample *v*, such as the market share of alternative *m* (i.e. modal market share), choice frequency, etc.

$\hat{s}^{e}{}_{v,m}$ is an aggregate outcome measure in sample *v*, such as the market share of alternative *m*, predicted from model estimated on sample *e*.

$\hat{P}\left(y^{e}_{n_v,m}\right)$ is the predicted probability that individual *n* chooses alternative *m,* predicted from model estimated on sample *e*.

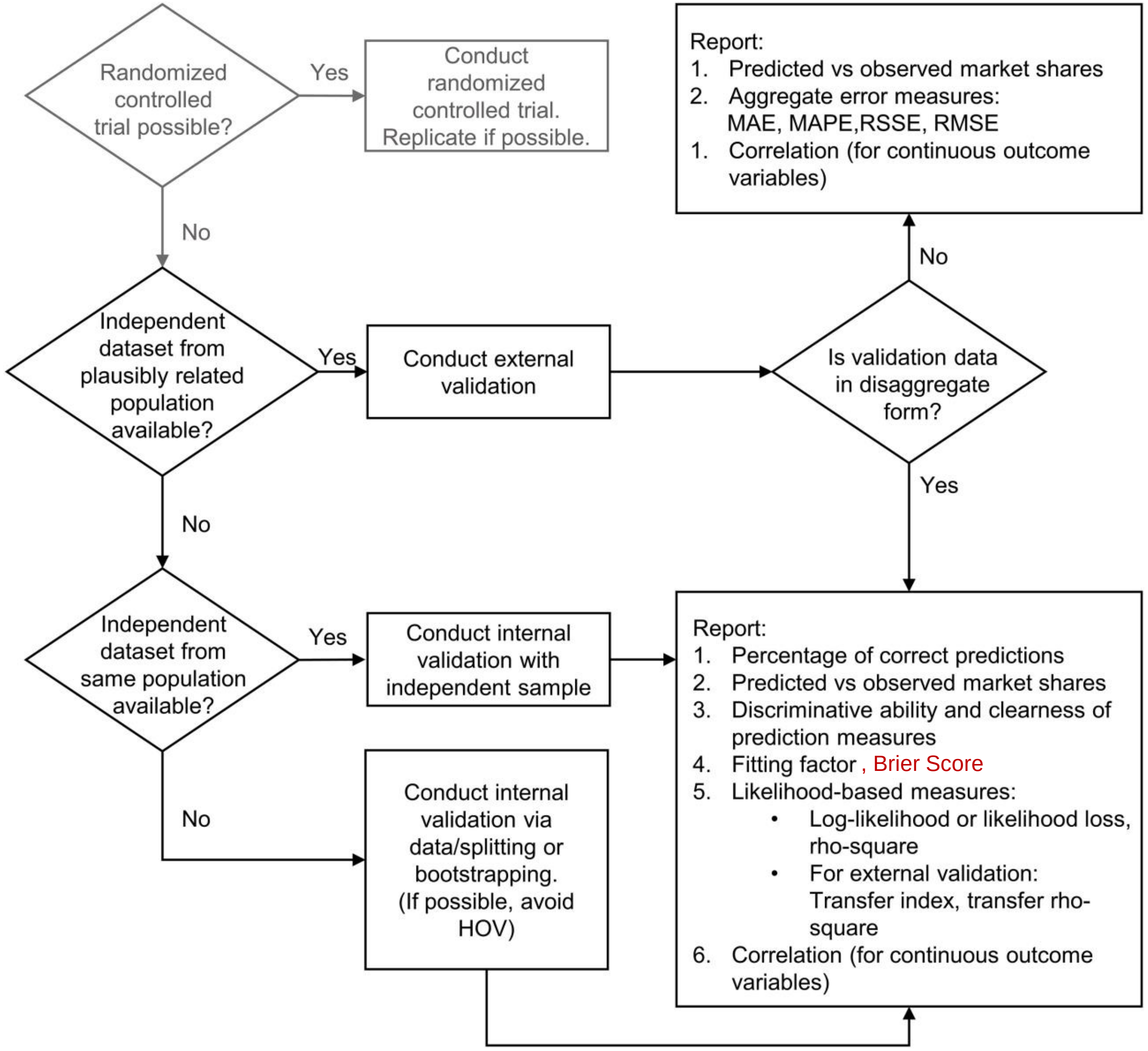
y_{nm} is the actual outcome variable valued 0 or 1.

Appendix: Definition of model validation performance measures reported in the literature

Parady, Ory & Walker (2021)

Index	Type		Formula	Notes
Log-likelihood 対数尤度	LL	Relative	$LL_v(\hat{\boldsymbol{\beta}}^e)$	$LL_{v,r}(\hat{\boldsymbol{\beta}}^e)$ is log-likelihood of the model estimated on data *e* applied to the validation data v_r.
Log-likelihood loss 対数尤度損失	LLL	Absolute	$\frac{1}{\mathrm{R}}\sum_r -\frac{1}{N_{v,r}}\sum_{n_{v,r}} LL_{v,r}(\hat{\boldsymbol{\beta}}^e)$ $\forall\, 1 \le r \le R$	$N_{v,r}$ is the size of the validation (holdout) sample r, and R is number of validation samples generated.
Rho-square σ^2	RHOSQ	Absolute	$\rho^2 = 1 - \frac{LL_v(\hat{\boldsymbol{\beta}}^e)}{LL_v(\mathbf{0})}$	$LL_v(\mathbf{0})$ is log-likelihood of the model when all parameters are zero for data *v*. $LL_v(\hat{\boldsymbol{\beta}}^v)$ is the likelihood of the model estimated on the validation data *v*.
Transfer rho-square 移転 σ^2	T-RHOSQ	Relative	$\rho^2_{transfer} = 1 - \frac{LL_v(\hat{\boldsymbol{\beta}}^e)}{LL_v(\boldsymbol{MS^v})}$	
Transfer index 移転指標	TI	Pass/Fail	$\frac{LL_v(\hat{\boldsymbol{\beta}}^e) - LL_v(\boldsymbol{MS^v})}{LL_v(\hat{\boldsymbol{\beta}}^v) - LL_v(\boldsymbol{MS^v})}$	$LL_v(\boldsymbol{MS^v})$ is a base model estimated on validation data *v* (i.e. market share model.)
Transferability test statistic 移転性検定統計量	TTS	Relative	$-2\left(LL_v(\hat{\boldsymbol{\beta}}^v) - LL_v(\hat{\boldsymbol{\beta}}^e)\right)$	ρ^2_{local} is the local rho-square of the model.
χ^2 **test**	CHISQ	Pass/Fail	$\sum_{m=1}^{M} \frac{\left(f_m - E\left(f^e_{v,m}\right)\right)^2}{E\left(f^e_{v,m}\right)}$	f_m is the observed choice frequency of alternative m in sample *v*, and $E\left(f^e_{v,m}\right)$ is the expected choice frequency predicted from model estimated on sample *e*.

Appendix: Validation and reporting practices in the transportation academic literature



Heuristic to select validation method given available resources and recommended performance measures to report

Appendix: Validation and reporting practices in the transportation academic literature

Table 4
Predictive accuracy performance measures reported in the literature by frequency.

Performance measure	Abbrv.	Frequency	Percentage
Log-likelihood/log-likelihood loss	LL/LLL	19	46.3%
Percentage of correct predictions or First Preference Recovery	FPR	10	24.4%
Predicted vs observed market outcomes	PVO	10	24.4%
Mean absolute error	MAE	6	14.6%
Root mean square error	RMSE	4	9.8%
Error/Percentage error/Absolute percentage error	E/PE/APE	3	7.3%
Rho-Square	RHOSQ	3	7.3%
Transfer index	TI	2	4.9%
% clearly right (t)	%CR	1	2.4%
Brier Score	BS	1	2.4%
Chi-square	CHISQ	1	2.4%
Concordance index	C	1	2.4%
Correlation	CORR	1	2.4%
Fitting factor	FF	1	2.4%
Mean absolute percentage error	MAPE	1	2.4%
Sum of square error	SSE	1	2.4%
Transferability test statistic	TTS	1	2.4%
All other measures specified in Table 1	–	0	0%
Other measures not specified in Table 1	–	3	7.3%

Very similar measures are reported jointly.