Eggers, Felix and John R. Hauser (2024), "Empirical Generalizations about Predictive Accuracy in Choice Experiments," *Advances in Consumer Research*, 52, forthcoming.

Empirical Generalizations about Predictive Accuracy in Choice Experiments

Introduction

The validity of choice experiments, i.e., conjoint analysis or discrete choice experiments, is typically assessed by comparing predicted and observed data. The most common validation metric is the hit rate of correct predictions, i.e., comparing for each consumer whether his or her predicted choice matches the observed choice in a holdout set or actual market choice. The range of hit rates can be determined by test-retest reliability as an upper limit (i.e., a consumer's consistency across repeated identical tasks) and a random selection among options as a lower limit. However, within this range no specific benchmarks exist about acceptable absolute hit rates and what determines the validity of choice models. This paper conducts a meta-analysis of predictive accuracy in choice experiments and aims to make empirical generalizations about hit rates that can be expected and create a better understanding of determinants of prediction accuracy and their underlying theories.

Conceptual model

As determinants of predictive validity, we include variables related to the research context (e.g., durables or services, novelty of the category), experimental design (e.g., number of alternatives per set, number of levels per attribute), estimation (e.g., number of choice sets, sample size), or craft of the implementation (e.g., using images vs. text, incentive alignment). These aspects can affect the predictive validity a) from a statistical point of view and b) due to behavioral processes. These processes can reinforce each other or counteract. For example, the larger the number of attributes the more parameters have to be estimated leading to potential prediction error. Behaviorally, more attributes can overwhelm respondents, making it difficult to identify preferred options, negatively impacting prediction accuracy. Conversely, more choice sets statistically imply more reliable estimates and better predictions, but can lead to decision fatigue and inconsistent choices, lowering accuracy.

Methodology

We conduct the meta-analysis on holdout hit rates in choice-based conjoint studies. Specifically, we identify all studies published in major marketing-related journals: Journal of Marketing Research (JMR), Journal of Marketing (JM), Marketing Science (MKSC), Management Science (MGMTSC), Journal of Consumer Research (JCR), Journal of the Academy of Marketing Science (JAMS), Journal of Product Innovation Management (JPIM), Quantitative Marketing and Economics (QME), International Journal of Research in Marketing (IJRM), and Marketing Letters (ML). We include all studies that were published until the year 2024 that mention "choice experiment" or "conjoint" and "hit rate" or "hit rates" in their text. Overall, we identified 184 observations nested in 86 studies. Our analysis uses a mixed-effects model that accounts for the fact that the hit rates are clustered within studies.

Results

Across all observations, we see a wide range of hit rates that vary between 7% and 88%. A main negative driver of the hit rate magnitude is the number of alternatives in the holdout set. Adjusting for it by measuring the relative improvement over a chance model shows an average factor of about 3 and a normalized improvement of about 40%. Predictive

accuracy is also affected to a large extent by the number of attributes (negative effect) and incentive alignment (positive effect).

Counterintuitively, the realism of the stimuli does not increase holdout prediction significantly. We find the same effect in Hauser, Eggers, and Selove (2019), in part because hit rate often measures internal validity rather than external validity. This warrants a further investigation into whether the appropriate criterion in choice experiments is internal or external validity. Although holdout sets are not used for the estimation, they are typically not distinguishable from regular choice tasks. If consumers adjust to the choice procedure and learn about their preferences in context of the choice experiment, which may nor may not represent real-world choices, they can make consistent choices even among less realistic, text-based alternatives, leading to no negative effect on hit rates in holdout sets. However, when holdout tasks mimic actual marketplace choices as a form of external validity, using realistic stimuli does have a significant positive effect on external validity (Hauser, Eggers, and Selove 2019).

Summary

Hit rates are the most prominent measure for predictive validity due to its ease of calculation and interpretation. This study presents a conceptual framework and empirical study of determinants of predictive validity as measured by hit rates. The results provide benchmarks to evaluate expected hit rates in published studies.

However, the underlying choice models are probabilistic in nature such that they cannot predict with certainty. Hit rates neglect this probabilistic nature and are insensitive to the scale of the estimates. For example, in a two-alternative choice context in which a logit model would predict choice probabilities of 0.51 and 0.49 the deterministic first-choice model that is required to calculate hit rates would predict that the first alternative would be chosen. Alternative validation criteria exist that consider choice uncertainty in the validation (e.g., Hauser 1978) but there is no consistency in reporting these measures.

Apart from reporting additional validity measures in studies, it would be desirable that studies also implement and report measures of test-retest reliability and attempt to compare internal and external validity. Knowing the choice consistency would allow to dissect the underlying behavioral and statistical effects better, which, in turn, provides a more reliable basis for making valid marketplace predictions.