

Imperfect Attention in Discrete Choice Experiments*

Sen Zeng[†]

May 15, 2026

Abstract

Discrete choice experiments (DCEs) are widely used to study policy when the policy is hypothetical or observational variation is limited, but their interpretation depends on whether respondents process the experimentally varied attributes that identify preferences. This paper develops a measurement-augmented latent-attention framework that separates preference heterogeneity from heterogeneity in attribute processing. The empirical comparison considers a benchmark full-attention mixed logit, a parsimonious parametric latent-attention specification, and a more flexible specification that places a shallow neural network in the prior over respondent-level attention states. The empirical implementation combines panel choice data with noisy post-experiment recall measures and estimates the model by generalized expectation-maximization (EM). I apply the framework to a menthol tobacco regulation DCE in which adult menthol smokers choose among cigarettes, e-cigarettes, and quitting under alternative price and legality scenarios. Both attention models fit the estimation sample better than the benchmark mixed logit. In respondent-level ex ante validation, however, the benchmark remains slightly better on held-out predictive metrics, while the attention models improve respondent-level later-task prediction once the model conditions on respondents' early held-out choices. The estimated conditional disutility of illegal menthol markets is substantially larger once incomplete processing is modeled explicitly. The counterfactual implications are economically meaningful: relative to the benchmark, the latent-attention models predict less quitting and greater persistence of menthol demand under prohibition. The results suggest that imperfect attention is not a secondary survey artifact. In policy-oriented DCEs, it can materially alter both structural interpretation and policy counterfactuals.

*I thank Donald Kenkel and Alan Mathios for providing the data and comments. Thanks to Levon Barseghyan, Francesca Molinari, and Colleen Carey for helpful feedback and discussions. All errors are my own.

[†]Department of Economics, Cornell University. E-mail: sz626@cornell.edu

1 Introduction

Discrete choice experiments (DCEs) have become a standard tool for empirical work in applied microeconomics, environmental economics, health economics, transportation, and marketing. They are especially attractive when researchers care about policies or product configurations that have not yet generated credible market data. The theoretical foundation for this enterprise is that properly designed survey questions can elicit preference information that is economically meaningful, particularly when respondents understand that the exercise is consequential for policy or implementation ([Carson and Groves, 2007](#); [Vossler and Watson, 2013](#)).

Yet the external validity of a DCE depends on more than randomization of attributes and incentive-compatible framing. A respondent may face the experimental menu and still not fully process what is being varied. In practice, hypothetical policy choice tasks are often cognitively dense. Respondents may ignore some attributes altogether, pay attention only to familiar products, or discount the implications of policy attributes that are difficult to map into concrete consequences. Once this possibility is admitted, standard DCE estimates become harder to interpret. Muted choice responses can reflect weak tastes, but they can also reflect incomplete processing of the attributes used to identify those tastes.

This distinction matters because standard choice data conflate utility and perception. As [Caplin \(2016\)](#) emphasizes, data on observed choices alone are generally insufficient when attention is endogenous. Related work in the economics of attention reaches a similar conclusion from several directions. Rational inattention models show that discrete choices emerge from a joint optimization over actions and information acquisition ([Caplin and Dean, 2015](#); [Matějka and McKay, 2015](#)). Salience models imply that decision weights depend on which attributes stand out in the local choice context ([Bordalo et al., 2013](#)). Sparse cognition models imply that decision makers simplify high-dimensional environments by focusing on a subset of first-order features ([Gabaix, 2014](#)). In a DCE, each of these mechanisms can generate the same empirical symptom: the coefficient on a policy attribute is not purely a taste parameter, but a taste parameter filtered through an endogenous processing weight.

This paper develops a measurement-augmented latent-attention framework for DCEs designed to separate, as far as the data allow, heterogeneity in preferences from heterogeneity in attribute processing. The empirical comparison is organized around three related specifications. The benchmark model is a standard full-attention mixed logit. The first attention-adjusted specification imposes a parsimonious prior over respondent-level attention states. The second keeps the same structural utility and measurement blocks but replaces the low-dimensional prior index with a shallow neural network. In both cases, noisy post-experiment recall responses enter as auxiliary measurement of the latent attention state, and estimation proceeds by a generalized expectation-maximization

(EM) algorithm.

The empirical application is a DCE conducted to study the potential prohibition of menthol cigarettes. Respondents repeatedly choose among non-menthol cigarettes, menthol cigarettes, tobacco-flavored e-cigarettes, menthol-flavored e-cigarettes, and quitting. Prices vary across combustible and e-cigarette products, and the legality of menthol products varies from legal availability to illegal availability through retail or street channels. This application is substantively important in its own right, but it also provides a useful methodological setting because the task involves both familiar products and cognitively demanding policy attributes, precisely the environment in which imperfect attention is likely to matter.

The descriptive evidence from the sample data points to selective processing. A nontrivial fraction of respondents make inertia choices, selecting the same option in all choice tasks. In addition, a post-DCE attention check shows that respondents are much more likely to correctly report variation in menthol cigarette price and legality than variation in less familiar products. These patterns are not conclusive on their own, strong tastes and habit persistence may also generate them, but they are difficult to reconcile with the view that every respondent pays equal attention to every experimentally varied attribute.

The model comparison yields four main findings. First, both latent-attention models improve substantially on the benchmark in in-sample fit. The machine-learning prior delivers the highest full-sample choice log likelihood and the lowest AIC, while the more parsimonious parametric prior delivers the lowest BIC. Second, modeling imperfect attention materially changes the structural interpretation of the experiment: relative to the benchmark, the attention models imply more negative conditional price and legality coefficients and less need to attribute muted responses to taste heterogeneity alone. Third, the out-of-sample evidence does not support a simple "more flexible is better" story. In respondent-level ex ante validation, the benchmark mixed logit remains slightly better on held-out predictive metrics. After conditioning on each held-out respondent's early choices, however, the attention models become more competitive and improve respondent-level later-task prediction. Finally, the counterfactual implications remain economically important: under menthol prohibition, the benchmark predicts more quitting and less persistence of menthol demand than either attention-adjusted model.

The paper proceeds as follows. Section 2 reviews the relevant literature. Section 3 presents the general latent-attention framework and the generalized EM estimator. Section 4 describes the menthol DCE and the auxiliary measurement data. Section 5 reports the structural estimates, validation results, and counterfactual policy simulations. Section 6 concludes with a discussion.

2 Related Literature

2.1 Stated-preference validity and the economics of attention

A first literature concerns the conditions under which survey-based preference data can be given an economic interpretation. [Carson and Groves \(2007\)](#) provide a foundational argument that preference questions are informative when respondents perceive them to be consequential. [Vossler and Watson \(2013\)](#) provide field evidence consistent with the importance of consequentiality for the validity of stated preference responses. This literature is often interpreted as asking whether respondents take the survey seriously. The present paper asks a different but complementary question: even if respondents do take the survey seriously, do they process the experimentally varied attributes in the way the researcher assumes?

That question connects directly to the economics of attention. [Caplin \(2016\)](#) reviews a broad literature arguing that choice data alone conflate utilities and perception, making standard revealed preference tools incomplete when attention is endogenous. [Caplin and Dean \(2015\)](#) develop revealed preference tests for costly information acquisition and rational inattention, formalizing the idea that apparent decision mistakes may instead reflect optimal limits on information processing. [Matějka and McKay \(2015\)](#) show that multinomial logit choice probabilities can be derived from rational inattention with Shannon information costs, thereby tying classical discrete choice tools to information frictions. [Bordalo et al. \(2013\)](#) show that consumers place disproportionate weight on salient attributes, while [Gabaix \(2014\)](#) develops a tractable sparse-cognition model in which agents simplify complex environments by focusing on first-order variables. Taken together, these papers suggest that incomplete or uneven processing is not an ad hoc survey pathology; it is an economically meaningful behavioral margin.

Recent work by [Abaluck et al. \(2020\)](#) shows how preferences can be recovered when consumers may be imperfectly informed because some attributes are hidden and consumers search. Their framework can be used to test full information and to study information counterfactuals and welfare under imperfect information. The present paper is complementary: rather than identify preferences from search-based restrictions alone, I combine randomized DCE variation with noisy post-task recall measures to estimate respondent-level heterogeneity in attribute processing inside a mixed logit demand system.

The contribution of this paper relative to the attention literature is empirical and methodological rather than theoretical. This paper does not attempt to recover a full rational inattention model from DCE data. Instead, I use the literature's central insight, that choice coefficients combine tastes and processing weights, to motivate a flexible empirical decomposition in survey choice settings.

2.2 Attribute non-attendance in discrete choice experiment

A second literature studies attribute non-attendance (ANA) directly in stated choice data. Early contributions documented that respondents often ignore one or more attributes when evaluating choice tasks and that accounting for this behavior can materially affect welfare measures and willingness-to-pay estimates. [Scarpa et al. \(2010\)](#) show that monitoring attribute attendance matters in a nonmarket valuation context. [Hole \(2011\)](#) develops a discrete choice model with endogenous attribute attendance and shows that it can outperform standard logit models while delivering different economic valuations.

Subsequent work pushed the literature in two directions. One branch compares stated and inferred attendance measures. [Hole et al. \(2013\)](#) estimate a mixed endogenous attendance model in which attendance probabilities depend on respondents' stated non-attendance and find that post-experiment stated ANA is informative about actual behavior. Another branch focuses on richer structures for heterogeneity in both tastes and ANA. [Collins et al. \(2013\)](#) emphasize that ANA and preference heterogeneity can be correlated and discuss important identification issues in generalized random-parameters ANA models. [Lew and Whitehead \(2020\)](#) review the ANA literature from the perspective of information processing and argue that ANA is best understood as part of a broader cognitive strategy rather than as a simple binary mistake.

This paper builds on that literature but shifts the emphasis. Rather than treating ANA as a standalone correction to willingness-to-pay estimates, I place the benchmark mixed logit, a parsimonious latent-attention model, and a more flexible latent-attention model inside one common measurement-augmented framework and compare them on fit, validation, and counterfactual policy implications.

2.3 Hybrid choice model, latent variables, and machine learning

A third literature introduces latent variables and richer psychological structure into discrete choice models. [Ben-Akiva et al. \(2002\)](#) define hybrid choice models that combine observed covariates, latent psychological constructs, and flexible disturbances within a unified framework. [Temme et al. \(2008\)](#) show how latent variables can be incorporated into discrete choice models using simultaneous estimation methods. These approaches preserve the structural interpretation of utility while allowing the analyst to capture attitudes, values, and perception-related variables that are not directly observed in market data. The present paper is close in spirit but more narrowly focused. The latent object is not a broad attitude or belief system; it is a respondent-level vector of processing indicators governing whether key price and legality attributes enter utility. The auxiliary data are noisy recall responses collected after the DCE. That is enough to create leverage on attention without requiring eye tracking, clickstream records, or richer process data.

Recent work on machine learning in choice modeling asks how flexible predictors can complement structural demand systems when nonlinearities or interactions are important (van Cranenburgh et al., 2022). In the ANA context, Díaz et al. (2023) show that supervised learning can be useful for respondent-level attendance classification. My use of machine learning is more disciplined than a standalone classifier. I do not estimate a black-box choice predictor, and I do not use the recall variables as first-step labels outside the structural system. Instead, the neural network enters only in the prior over respondent-level latent attention states and is estimated jointly with the utility and measurement blocks inside the generalized EM algorithm. This preserves coefficient interpretation and counterfactual discipline while allowing more flexible nonlinearities in the determinants of attention.

3 The Model

This section develops a general framework for imperfect attention in discrete choice experiments. The empirical comparison considers a benchmark full-attention mixed logit, a parsimonious latent-attention model, and a more flexible specification in which the prior distribution of attention is modeled with a regularized nonlinear predictor.

3.1 General setup

Consider respondent $i = 1, 2, \dots, N$ choosing among J alternatives in each of T choice tasks. Let x_{ijkt} denote the vector of observed attributes k for alternative j in task t . It is useful to partition x_{ijkt} into two components. The first, c_{ijkt} , contains variables that enter utility regardless of attention. The second, z_{ijkt} , contains variables whose contribution depends on whether the respondent processes the corresponding attribute.

Let $\beta_i = (\gamma_i, \delta_i)$ denote respondent-specific tastes, where γ_i is associated with c_{ijkt} and δ_i is associated with z_{ijkt} . Let $A_i = (A_{i1}, A_{i2}, \dots, A_{ik})$ be a vector of latent attention indicators, with A_{ik} equal to one if respondent i attends to attribute k and zero otherwise. The general utility representation considered here is:

$$U_{ijt}(A_i, \beta_i) = \alpha_j + c'_{ijt}\gamma_i + (A_i \odot z_{ijt})'\delta_i + \varepsilon_{ijt} \quad (1)$$

where α_j is an alternative-specific constant, $\odot$ denotes elementwise multiplication, and ε_{ijt} is an idiosyncratic error term.

This representation is general enough to conceptually nest the main empirical cases considered here. If $A_{ik} = 1$ for all respondents and all attributes, the model collapses to a conventional full-attention random utility model. If A_{ik} is binary, the model coincides with a standard attribute

non-attendance specification. More generally, the same representation can be interpreted as a reduced-form approximation to richer models in which decision makers place unequal cognitive weight on different attributes.

Conditional on (A_i, β_i) , and assuming the usual type-I extreme-value distribution for ε_{ijt} , the probability that respondent i chooses alternative j in task t is multinomial logit:

$$P_{ijt}(A_i, \beta_i) = \frac{\exp\{\alpha_j + c'_{ijt}\gamma_i + (A_i \odot z_{ijt})'\delta_i\}}{\sum_{l=1}^J \exp\{\alpha_l + c'_{ilt}\gamma_i + (A_i \odot z_{ilt})'\delta_i\}} \quad (2)$$

Let y_i denote respondent i 's sequence of observed choices, the panel likelihood conditional on (A_i, β_i) is the product of these task-level choice probabilities across all choice occasions.

This formulation makes the identification problem transparent. The choice data reveal only the product of tastes and attention. A weak choice response to an attribute can reflect weak tastes, low attention, or both. Recovering attention separately from tastes therefore requires either auxiliary measurements, exclusion restrictions, or additional structure.

3.2 Benchmark model: full-attention mixed logit

The benchmark model imposes full attention by setting $A_i = \mathbf{1}_K$ for every respondent, where $\mathbf{1}_K$ denotes a K -dimensional vector of ones. Under this restriction, utility reduces to a standard mixed logit with respondent-specific random coefficients. Let $f_\beta(\beta_i|\theta)$ denote the mixing distribution for β_i . The respondent-level likelihood contribution is

$$L_i^B(\theta) = \int \prod_{t=1}^T P(y_{it}|A_i = \mathbf{1}_K, \beta_i) f(\beta_i|\theta) d\beta_i \quad (3)$$

This benchmark is useful for two reasons. First, it is flexible with respect to taste heterogeneity, and provides a familiar baseline against which attention-adjusted models can be compared. Second, it clarifies the substantive issue at stake. Under full attention, any attenuation in the observed response to an experimentally varied attribute must be absorbed by tastes or by residual error. If respondents selectively process some attributes, the benchmark risks interpreting attention heterogeneity as taste heterogeneity.

3.3 Measurement-augmented latent attention

To create leverage on attention, suppose we also observe auxiliary measurements $R_i = (R_{i1}, R_{i2}, \dots, R_{iK})$, where R_{ik} is informative about whether respondent i processed attribute k . These signals may be binary, ordinal, or categorical. They are not assumed to be literal observations of real-time cognition, rather, they are treated as noisy, reduced-form measurements of the underlying latent attention state.

Let η denote the parameters of the measurement system. A convenient specification assumes conditional independence across attention dimensions:

$$P(R_i|A_i; \eta) = \prod_{k=1}^K P(R_{ik}|A_{ik}; \eta_k) \quad (4)$$

When R_{ik} is categorical, the component likelihood can be written as a set of category probabilities indexed by the latent attention state. This formulation is flexible enough to accommodate recall questions, self-reported processing measures, or other indirect indicators collected after the choice tasks have been completed.

The auxiliary responses should be interpreted as reduced-form measurements rather than literal observations of cognition at the moment of choice. In many applications, the available indicators summarize whether respondents noticed or retained variation in the experimentally manipulated attributes. That information is imperfect, but it can still help distinguish weak tastes from incomplete processing.

A natural and substantively appealing normalization is monotonicity. Responses that are more consistent with attention should be weakly more likely when $A_{ik} = 1$ than when $A_{ik} = 0$. This restriction helps stabilize estimation and rules out label switching in the latent attention states.

3.4 Prior distribution of attention states

Let W_i denote observed variables that shift attention but are excluded from utility conditional on the attributes that appear in the choice problem. The latent attention vector is distributed according to

$$P(A_i = a|W_i; \psi), \quad a \in \{0, 1\}^K \quad (5)$$

The paper considers two specifications for this prior. The first is a parsimonious parametric prior. Let $\pi_{ik} = P(A_{ik} = 1|W_i; \psi)$. If the attention indicators are conditionally independent given W_i , the baseline prior can be written as the product of Bernoulli probabilities across attention dimensions. This baseline prior may be enriched by allowing an additional mass on the full-attention state. That extension is useful when the data suggest that a nontrivial share of respondents attend to all relevant attributes.

The second specification replaces the low-dimensional parametric index with a regularized nonlinear predictor. The role of machine learning is confined to the prior over latent attention. Utility remains structural, and the measurement system is unchanged. This division of labor preserves interpretability while allowing more flexible interactions and nonlinearities in the determinants of attention.

Combining the choice model, the attention prior, and the measurement system yields the

respondent-level likelihood

$$L_i(\theta, \psi, \eta) = \sum_a P(A_i = a | W_i; \psi) P(R_i | A_i = a; \eta) \int P(y_i | A_i = a, \beta_i) f_\beta(\beta_i | \theta) d\beta_i \quad (6)$$

In practice, the integral over β_i is approximated by simulation. Let $\beta_{ir}(\theta)$ denote the r th simulated draw for respondent i . Then the simulated likelihood is

$$\hat{L}_i(\theta, \psi, \eta) = \sum_a P(A_i = a | W_i; \psi) P(R_i | A_i = a; \eta)^\omega \left[\frac{1}{R} \sum_r P(y_i | A_i = a, \beta_{ir}(\theta)) \right] \quad (7)$$

When the auxiliary measurements are noisy, it is useful to allow their contribution to be down-weighted relative to the choice data. A convenient implementation raises the measurement term to a power ω in $(0, 1]$. The case $\omega = 1$ corresponds to the standard joint likelihood. Values below one retain the directional information in the measurement block while limiting the influence of noisy auxiliary responses.

3.5 Identification and estimation

The model is identified from several sources of variation. First, the experimental design must generate within-respondent variation in the attributes whose processing is in question. Without such variation, attention to a given attribute is not empirically meaningful. Second, the auxiliary measurements must be informative about the latent attention state. They need not be perfect. It is enough that they shift the posterior distribution of attention in a systematic way. Third, the variables W_i that enter the attention prior must be excluded from utility after conditioning on the experimentally varied attributes. This exclusion restriction is central. It need not be interpreted as a claim about deep psychological primitives. It is enough that the variables in W_i affect the likelihood that respondents process particular attributes, while their direct effect on utility is negligible once the choice attributes are included. Fourth, the usual scale normalization of multinomial logit remains in force. Utility is identified only up to location and scale.

Finally, the level of the latent attention object should match the level at which auxiliary information is observed. When auxiliary measurements summarize processing over the full sequence of tasks, respondent-level attention is the natural empirical object. Task-level attention is feasible only when the data include task-level process measures or when the analyst is willing to impose stronger dynamic structure.

The benchmark mixed logit is estimated by maximum simulated likelihood. The latent-attention models are estimated by a generalized EM algorithm because respondents' attribute-attention states are unobserved. At iteration q , let $\tau_{isr}^{(q)}$ denote the posterior responsibility attached to attention state

s and simulation draw r for respondent i . Up to normalization,

$$\tau_{isr}^{(q)} \propto p(s|W_i; \psi^{(q)}) \times P(R_i|s; \eta^{(q)})^\omega \times P(y_i|s, \beta_{ir}(\theta^{(q)})) \quad (8)$$

The E-step computes these posterior weights for every respondent, attention state, and simulation draw. The M-step then updates the parameter blocks in turn. The utility update chooses θ to maximize the expected complete-data log-likelihood for the choice block, holding the current posterior weights fixed. The measurement update chooses η to maximize the weighted measurement log-likelihood. With categorical measurement probabilities and Dirichlet smoothing, this step has a closed-form update. The prior update chooses ψ to maximize the weighted log prior, possibly with regularization when the prior is modeled flexibly. Figure 1 presents a description of the algorithm.

Because the utility and prior updates are solved numerically rather than analytically, the procedure is a generalized EM algorithm rather than an exact EM algorithm. In practice, convergence should be assessed using both the evolution of the objective function and the stability of economically relevant outputs, including transformed coefficients, predictive scores, and counterfactual market shares. Estimation follows a standard EM logic for latent-variable models (Dempster et al., 1977; McLachlan and Krishnan, 2008), adapted here to a mixed-logit setting with simulated integration over random coefficients (Train, 2009). Details on the empirical application’s implementation of initialization, simulation draws, attention measurement, tuning parameters and validation are reported in Appendix A.

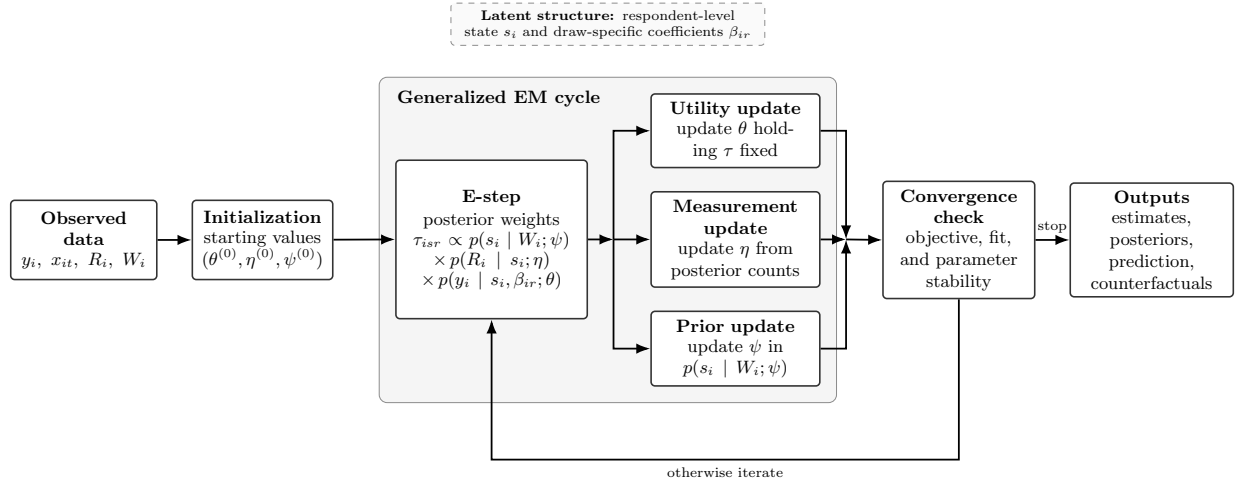


Figure 1: Generalized Expectation-Maximization Algorithm for a Latent Attention Model

4 Empirical Application

4.1 The menthol prohibition discrete choice experiment

The empirical application is a DCE designed to assess the consequences of prohibiting menthol cigarettes in the United States. [Kenkel et al. \(2025\)](#) provide the policy background and institutional motivation. The sample contains 639 adult menthol smokers who have smoked 100 or more cigarettes in their lifetime. Each respondent completed 12 choice tasks, yielding 7,668 respondent-task observations. In every task, the respondent chose among five alternatives: non-menthol cigarettes, menthol cigarettes, tobacco-flavored e-cigarettes, menthol-flavored e-cigarettes, or quitting smoking cigarettes not using e-cigarettes.

The price and legality attributes vary across the product alternatives. For combustible cigarettes, price is expressed relative to each respondent's own recent purchase price: $0.5p$, p , or $2p$, where p denotes the price paid for the respondent's last pack of cigarettes¹. For e-cigarettes, price takes the values \$2, \$4, or \$8. I treat \$4 as the approximate baseline e-cigarette price for a quantity roughly equivalent to a pack of cigarettes. Legal availability is especially important for the policy application. Non-menthol products remain legal. Menthol cigarettes and menthol-flavored e-cigarettes can instead be legal, prohibited but available under the counter and through online retailers, or prohibited and strictly enforced so that they are available only from illegal dealers such as street sellers. Table 1 shows the alternatives and variation of associated attributes presented in the DCE. In the DCE, each subject was presented 12 scenarios that each describes a choice set and was asked to make a choice after each scenario presentation.² An example of the choice scenario presentation in the DCE is shown in figure 2.

The design has several attractive features for studying imperfect attention. First, it combines attributes that are likely to be cognitively easy to process, such as own-product price, with attributes that require richer consequence mapping, such as whether an illegal product remains available through informal channels. Second, it mixes familiar and less familiar products. All respondents are adult menthol smokers, but not all have recent experience with non-menthol cigarettes or tobacco-flavored e-cigarettes. Third, the outside option of quitting provides an extensive-margin behavioral response that is highly policy relevant but potentially sensitive to how respondents interpret illegality and substitution possibilities.

¹The cost question asked in the survey is “What price did you pay for the last pack of cigarettes? Please report the cost after using discounts or coupons”.

²Given that the experimental design consists of 3 levels of menthol cigarette prices, 3 levels of non-menthol cigarette prices, 3 levels of menthol e-cigarette prices, 3 levels of non-menthol e-cigarette prices, 3 levels of menthol cigarette legality conditions, and 3 levels of menthol e-cigarette legality conditions, the number of possible combinations of attributes is 729. The design uses 12 scenarios to balance respondent burden against statistical efficiency for the effects of interest.

Table 1: Alternatives and associated attributes in the DCE

Alternatives	Attributes	
	Price	Legality
Non-menthol cigarettes	0.5p; p; 2p	legal
Menthol cigarettes	0.5p; p; 2p	legal
		Prohibited. Available under-the-counter and online from retailers who continue to sell prohibited menthol products
		Prohibited. Strictly enforced, only available from illegal dealers, e.g. street sellers
Tobacco flavored e-cigarettes	\$2; \$4; \$8	legal
Menthol flavored e-cigarettes	\$2; \$4; \$8	legal
		Prohibited. Available under-the-counter and online from retailers who continue to sell prohibited menthol products
		Prohibited. Strictly enforced, only available from illegal dealers, e.g. street sellers

Quitting

Notes: p represents the baseline price of cigarettes, which is measured by the price that an individual paid for last pack of cigarettes.

Here is another set of products that could be available to you.

Think about your immediate choice **today**. Here are the set of cigarettes and e-cigarettes available when you are shopping.

Please select one option **for your immediate choice today** from the choices below.

(If you want to see a larger version of the images, please click the magnifying glass below.)

	Non-Menthol Cigarettes	Menthol Cigarettes	Tobacco Flavored E-cigarettes	Menthol Flavored E-cigarettes	None
Product					I will quit smoking cigarettes and not use e-cigarettes
Price	\$10.00	\$5.00	\$8.00	\$4.00	
Legality	Legal	Legal	Legal	Prohibited. Strictly enforced. Only available from illegal dealers, e.g. street sellers	







Figure 2: An example choice task from the menthol DCE

4.2 Descriptive evidence on imperfect attention

After completing the experiment, respondents were asked to think back to the earlier choice tasks and indicate which product features had varied from question to question. The attention check covered the legality or availability of menthol cigarettes and menthol-flavored e-cigarettes and the prices of the product alternatives. This type of ex post recall question is not a perfect measure of real-time attention, but it has two important advantages. It is cheap to collect in standard online surveys, and it directly asks about the variation that identifies the structural coefficients in the DCE. In the model below, I treat them as noisy auxiliary measurement of attribute processing; in the descriptive analysis, they also provide a direct diagnostic of selective encoding.

Two descriptive patterns stand out. The first is inertia. Some respondents choose the same alternative in all 12 tasks. Because all respondents are menthol smokers, one possible explanation is strong pre-existing product attachment. But inertia can also be a symptom of simplified decision rules or low engagement. The raw distribution in table 2 suggests that both interpretations may be relevant. Menthol cigarettes account for 43.21 percent of all choices and 12.05 percent of respondents choose menthol cigarettes in every task. Quitting accounts for 22.10 percent of all choices and 4.54 percent of respondents choose quit in every task. Menthol e-cigarettes account for 17.75 percent of all choices and 2.03 percent of respondents choose menthol e-cigarettes in every task. Tobacco-flavored e-cigarettes and non-menthol cigarettes receive much smaller shares overall and very little all-task inertia.

Table 2: Choice patterns

Products	Used in Past Year	All Choices in DCE	Inertia Choices
Menthol cigarettes	100%	43.21%	12.05%
Tobacco flavored cigarettes		8.93%	0.16%
Menthol e-cigarettes	41.56%	17.75%	2.03%
Tobacco flavored e-cigarettes	18.92%	8.01%	0.31%
Quit	60.67%	22.10%	4.54%

Notes: For products used in past year, we define the fraction of choosing quit as those have tried quitting in the past year, the fractions of products do not sum to one because some subjects use multiple products at the same time (e.g., dual users of cigarettes and e-cigarettes). Inertia choice is defined as choosing the same alternative across 12 choice scenarios.

The second pattern is selective recall of attribute variation. Figure 3 shows that the share of respondents who correctly report that a given attribute varied is only 31% for non-menthol e-cigarette price, 40% for non-menthol cigarette price, 41% for menthol e-cigarette price, and 39% for menthol e-cigarette legality. By contrast, 59% correctly report variation in menthol cigarette price and 56% correctly report variation in menthol cigarette legality. In other words, even after twelve repeated choice tasks, attribute variation is remembered by only a minority or slim majority of respondents,

and recall is clearly stronger for the attributes of the familiar menthol cigarette product.

These descriptive facts are consistent with endogenous attention. Salient and familiar products appear to receive more cognitive weight. The patterns are not themselves proof of a particular attendance model, but they provide exactly the sort of diagnosis that motivates a structural comparison between full-attention and attention-adjusted specifications.

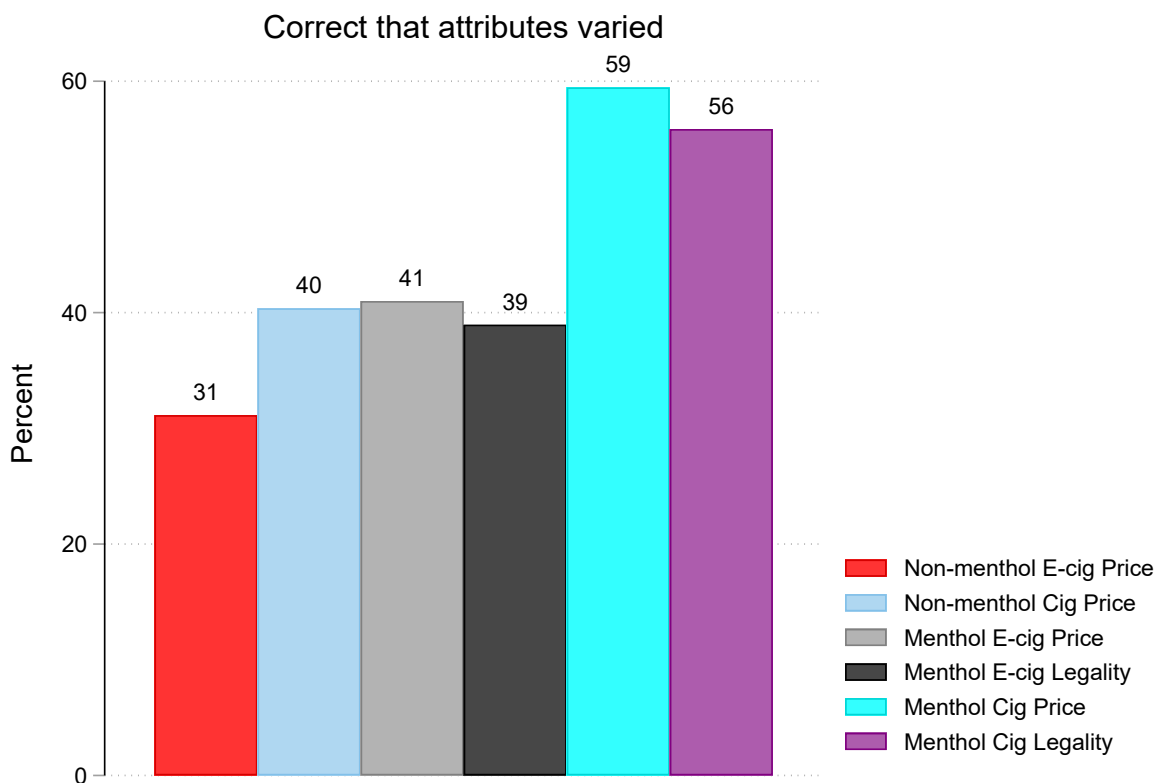


Figure 3: Percent of respondents correctly recalled attributes variation in the DCE

4.3 Estimation and Validation Design

All three models share the same utility specification. Quitting is the omitted base alternative, so utility includes four alternative-specific constants for the product options, one common price coefficient, and four legality indicators: menthol cigarettes retail-illegal, menthol cigarettes street-illegal, menthol e-cigarettes retail-illegal, and menthol e-cigarettes street-illegal. Non-menthol cigarettes and tobacco-flavored e-cigarettes remain legal in all scenarios and therefore carry no legality terms.

The benchmark mixed logit uses four normally distributed alternative-specific constants and five negative lognormal coefficients for price and legality. The latent-attention models retain the same

utility block but augment it with respondent-level latent attention over six dimensions: menthol-cigarette legality, menthol e-cigarette legality, non-menthol cigarette price, menthol cigarette price, tobacco-flavored e-cigarette price, and menthol-flavored e-cigarette price. The latent state therefore has six binary components and $2^6 = 64$ possible attention patterns.

The auxiliary measurement system is built from the post-experiment recall block. For each of the six attention dimensions, respondents are asked whether the relevant attribute varied across the tasks. The model specification recodes those responses into three categories—correct recall, ambiguous recall, and no recall—and imposes a monotonicity restriction so that responses more consistent with correct recall are weakly more likely under attention than under inattention. Because these measures summarize cumulative processing rather than task-specific cognition, the measurement block is down-weighted relative to the choice block with $\omega = 0.1$.

The respondent-level variables used in the attention prior are drawn from survey process measures: log time spent in the DCE, an indicator for completing the survey on a phone, and an indicator for multiple survey visits. In the parametric model, the prior contains dimension-specific intercepts, a shared slope vector for legality-attention dimensions, a shared slope vector for price-attention dimensions, and an additional full-attention mass parameter. In the flexible model, the same information enters a one-hidden-layer neural network. Both models are estimated jointly with the utility and measurement blocks by generalized EM.

Validation uses two respondent-level exercises. The first is 5-fold ex ante cross-validation: respondents are split into folds, the models are refit on the training respondents, and the fitted models predict the held-out respondents using only the prior over attention states. The second is an adaptive later-task exercise: within each held-out respondent, the model conditions on the first six tasks and predicts the remaining tasks. The ex ante exercise is the relevant “cold-start” test; the adaptive exercise asks whether the latent-attention structure becomes more useful after the researcher learns something about the respondent from her own choices.

5 Results

5.1 Structural estimates

Table 3 reports the structural estimates and the in-sample fit statistics for the benchmark mixed logit, the parametric latent-attention model, and the machine-learning latent-attention model. The first result is straightforward: both attention models fit the estimation sample better than the benchmark. The benchmark attains a choice log-likelihood of -6736, whereas the parametric and machine-learning attention models improve this to -6607 and -6588, respectively. By AIC, the machine-learning prior performs best, while by BIC the more parsimonious parametric prior

performs best. The in-sample fit comparison therefore favors explicit attention modeling, but it also suggests that most of the economically relevant regularities in attention are fairly low dimensional in this application.

The benchmark coefficients are economically sensible. The alternative-specific constant for menthol cigarettes is large and positive, reflecting the baseline attachment of this sample to menthol combustible products. Menthol-flavored e-cigarettes also receive a positive constant, whereas tobacco-flavored e-cigarettes receive a large negative constant. The mean price coefficient is negative, and illegal availability lowers utility for menthol products. The street-illegal coefficients are more negative than the retail-illegal coefficients, which is what one would expect if a strictly illicit market is less attractive than a market in which prohibited products remain available through under-the-counter or online retail channels.

The attention-adjusted models change the interpretation of those utility coefficients in systematic ways. Start with price. The mean price coefficient becomes more negative once attention is modeled explicitly, moving from -0.383 in the benchmark to -0.439 in the parametric attention model and -0.440 in the machine-learning model. At the same time, the reported standard deviation of the price coefficient falls from 0.651 in the benchmark to 0.484 and 0.487 in the two attention models. This is an important pattern. Under the full-attention benchmark, attenuation in choice responses must be absorbed by either weaker tastes or greater taste heterogeneity. Once the model allows some respondents to process price imperfectly, less of that attenuation is forced into the random-coefficients distribution.

The same logic is even clearer for legality. Relative to the benchmark, both attention models imply substantially larger disutility from illegal menthol markets. For menthol cigarettes, the illegal-retail coefficient shifts from -1.543 in the benchmark to -3.015 and -3.055 in the parametric and machine-learning models, respectively, and the illegal-street coefficient shifts from -2.149 to -4.205 and -4.229. For menthol-flavored e-cigarettes, the illegal-retail coefficient shifts from -1.533 to -3.076 and -3.070, while the illegal-street coefficient shifts from -2.536 to -4.690 and -4.703. In economic terms, the benchmark appears to understate how strongly respondents dislike prohibition environments conditional on actually processing the legality information.

These shifts do not mean that the attention models simply make respondents more policy-sensitive across the board. The alternative-specific constants move at the same time. The constant for menthol cigarettes rises from 4.472 in the benchmark to 5.608 and 5.627 in the attention models, while the constant for menthol-flavored e-cigarettes rises more modestly from 1.728 to 2.015 and 2.020. Tobacco-flavored e-cigarettes become slightly less unattractive, and the non-menthol cigarette constant remains close to zero and slightly negative. The point is not that every coefficient moves in the same direction. The point is that the attention models recover a different decomposition of behavior: stronger latent attachment to menthol cigarettes, stronger conditional disutility from price

and illegality, and less need to explain muted responses by loading everything into taste heterogeneity alone.

That decomposition is the economic content of the latent-attention framework. In the benchmark mixed logit, the same observed behavior is interpreted entirely through tastes. In the attention models, part of the attenuation in observed choice responses is reallocated to incomplete or selective processing of experimentally varied attributes. The resulting estimates therefore differ not just in magnitude but in interpretation. The benchmark conflates preference heterogeneity with heterogeneity in attribute processing, whereas the attention-adjusted models separate those two margins as far as the data allow.

Finally, the comparison between the two attention models is itself informative. The machine-learning prior buys a small additional improvement in log-likelihood and AIC, but the parametric model remains highly competitive once model complexity is penalized. That pattern suggests that the main attention regularities in this DCE are not especially high dimensional. A flexible prior is still useful as a robustness check and as a guard against functional-form misspecification, but the substantive conclusions do not depend on a black-box attention equation.

Although the latent-attention formulation contains full attention as a boundary special case, this does not by itself justify regular likelihood-ratio inference. I therefore treat in-sample log-likelihood and information criteria as descriptive fit summaries rather than as the basis for formal model-selection tests. Table 4 reports pairwise Vuong test comparisons based on respondent-level contributions to the common choice likelihood. Both attention models significantly outperform the benchmark full-attention mixed logit: relative to the benchmark, the parametric attention model is favored by the unadjusted, AIC-adjusted, and BIC-adjusted Vuong statistics, and the same is true for the ML attention model. Comparing the two attention models directly, the ML attention model has the better unadjusted fit, but that advantage is modest once model complexity is taken into account. The AIC-adjusted comparison no longer rejects equal fit, while the BIC-adjusted statistic favors the more parsimonious parametric attention model. Taken together, the Vuong results support the conclusion that allowing for latent attention materially improves respondent-level choice fit relative to the benchmark, while the incremental gain from replacing the parametric prior with a more flexible ML prior is limited after penalizing complexity.

5.2 Predictive validation

The validation results are intentionally reported as a horse race between cold-start and adaptive prediction. Figure 4 reports 5-fold benchmark-relative out-of-sample prediction using two conceptually distinct criteria: held-out pseudo- R^2 , which summarizes respondent-level predictive fit, and held-out share RMSE, which summarizes the accuracy of aggregate predicted market shares. This exercise

Table 3: Estimation results

	Benchmark		Parametric attention		ML attention	
	Mean	SD	Mean	SD	Mean	SD
ASC (Non-menthol cigarettes)	-0.167 (0.221)	3.112*** (0.222)	-0.189 (0.173)	3.560*** (0.172)	-0.178 (0.176)	3.565*** (0.176)
ASC (Menthol cigarettes)	4.472*** (0.165)	3.096*** (0.204)	5.608*** (0.147)	3.620*** (0.140)	5.627*** (0.148)	3.632*** (0.140)
ASC (Tobacco-flavored e-cigarettes)	-2.055*** (0.322)	3.912*** (0.277)	-1.956*** (0.211)	3.781*** (0.187)	-1.936*** (0.211)	3.782*** (0.186)
ASC (Menthol-flavored e-cigarettes)	1.728*** (0.152)	3.303*** (0.183)	2.015*** (0.128)	3.804*** (0.165)	2.020*** (0.129)	3.819*** (0.163)
Price (\$)	-0.383*** (0.022)	0.651*** (0.092)	-0.439*** (0.018)	0.484*** (0.046)	-0.440*** (0.018)	0.487*** (0.046)
Illegal Retail Market for Menthol Cigarettes	-1.543*** (0.119)	1.185*** (0.312)	-3.015*** (0.270)	1.577*** (0.433)	-3.055*** (0.267)	1.587*** (0.431)
Illegal Street Market for Menthol Cigarettes	-2.149*** (0.174)	1.665*** (0.613)	-4.205*** (0.277)	1.569*** (0.412)	-4.229*** (0.277)	1.584*** (0.418)
Illegal Retail Market for Menthol E-cigarettes	-1.533*** (0.119)	0.478*** (0.194)	-3.076*** (0.225)	0.783*** (0.178)	-3.070*** (0.219)	0.778*** (0.179)
Illegal Street Market for Menthol E-cigarettes	-2.536*** (0.259)	2.579*** (0.805)	-4.690*** (0.246)	1.078*** (0.774)	-4.703*** (0.251)	1.081*** (0.722)
Log-likelihood	-6736		-6607		-6588	
AIC	13509		13276		13269	
BIC	13589		13414		13479	
Observations	7668		7668		7668	

Notes: Entries report transformed coefficient means and standard deviations. For normal coefficients (the four alternative-specific constants), the reported mean and standard deviation coincide with the estimated location and scale parameters. For the negative-lognormal coefficients (price and legality attributes), the reported mean and standard deviation are transformations of the underlying normal parameters. In the attention models, the price and legality coefficients are conditional on respondents processing the corresponding attribute; population-average behavioral responses therefore depend jointly on those coefficients and the estimated attention probabilities. Benchmark standard errors are analytic simulated-maximum-likelihood standard errors. Attention-model standard errors are respondent-bootstrap standard errors computed by re-estimating the full model on bootstrap resamples. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Pairwise Vuong Test Comparisons

Comparison	Unadjusted z	Favored	AIC-adjusted z	Favored	BIC-adjusted z	Favored
Benchmark vs. Parametric attention	-7.78***	Parametric attention	-7.22***	Parametric attention	-5.98***	Parametric attention
Benchmark vs. ML attention	-8.24***	ML attention	-7.06***	ML attention	-4.42***	ML attention
Parametric attention vs. ML attention	-3.87***	ML attention	-0.86	Neither	5.85***	Parametric attention

Notes: Each row reports a pairwise Vuong comparison using respondent-level contributions from the common choice likelihood. Negative z -statistics favor the second model listed in the comparison; positive z -statistics favor the first. The adjusted statistics apply AIC and BIC complexity penalties, respectively. Because the attention models are compared on the common choice likelihood, the table abstracts from the weighted recall-measurement block used during estimation. *** $p < 0.01$.

separates the key predictive question into a respondent-level dimension and an aggregate-share dimension.

Note that better respondent-level fit does not mechanically imply better aggregate fit. In the ex ante prediction exercise, both attention models perform worse than the benchmark on held-out pseudo- R^2 and on held-out share RMSE, indicating that the full-attention mixed logit remains the strongest cold-start predictor. By contrast, once the models are allowed to condition on respondents' early choices, both attention models improve respondent-level predictive fit relative to the benchmark, as shown by the higher later-task pseudo- R^2 . At the same time, both attention models continue to have slightly worse held-out share RMSE than the benchmark. This pattern is substantively useful: it shows that the attention models help explain respondent-specific heterogeneity once some individual choice history is observed, but that this gain does not automatically translate into more accurate aggregate market-share prediction. The right comparison is therefore not “which model is uniformly best?” but rather “what kind of prediction problem is the model solving?” For cold-start respondent-level prediction, the benchmark is hard to beat. For conditional prediction after the analyst has learned something about the respondent, the latent-attention structure becomes more useful.

This distinction is what one should expect from a measurement-augmented latent-state model. Attention is only partly predictable from coarse respondent-level prior features. Once the model can condition on a respondent's own early choices, it learns more about both tastes and processing and can use that information to improve later-task prediction. The validation results therefore support a limited but credible claim: explicit attention modeling is valuable, but primarily for structural interpretation and conditional prediction, not as a guaranteed improvement in pure ex ante holdout performance.

5.3 Counterfactual policy simulations

The strongest economic case for modeling attention comes from the counterfactual simulations in Figure 5. Two qualitative patterns are robust. First, relative to the benchmark, the attention-adjusted models place more baseline weight on menthol cigarette demand and less on quitting, a difference that is already visible under the status quo. Second, as menthol products become illegal, the benchmark predicts a stronger movement into quitting than either attention model. The difference is largest when both menthol cigarettes and menthol-flavored e-cigarettes are illegal and available only through illicit channels.

This result is economically important because it does not follow mechanically from the coefficient table. One might expect the larger negative legality coefficients in the attention models to imply more quitting, not less. But the counterfactuals reflect the entire system. Once attention is modeled

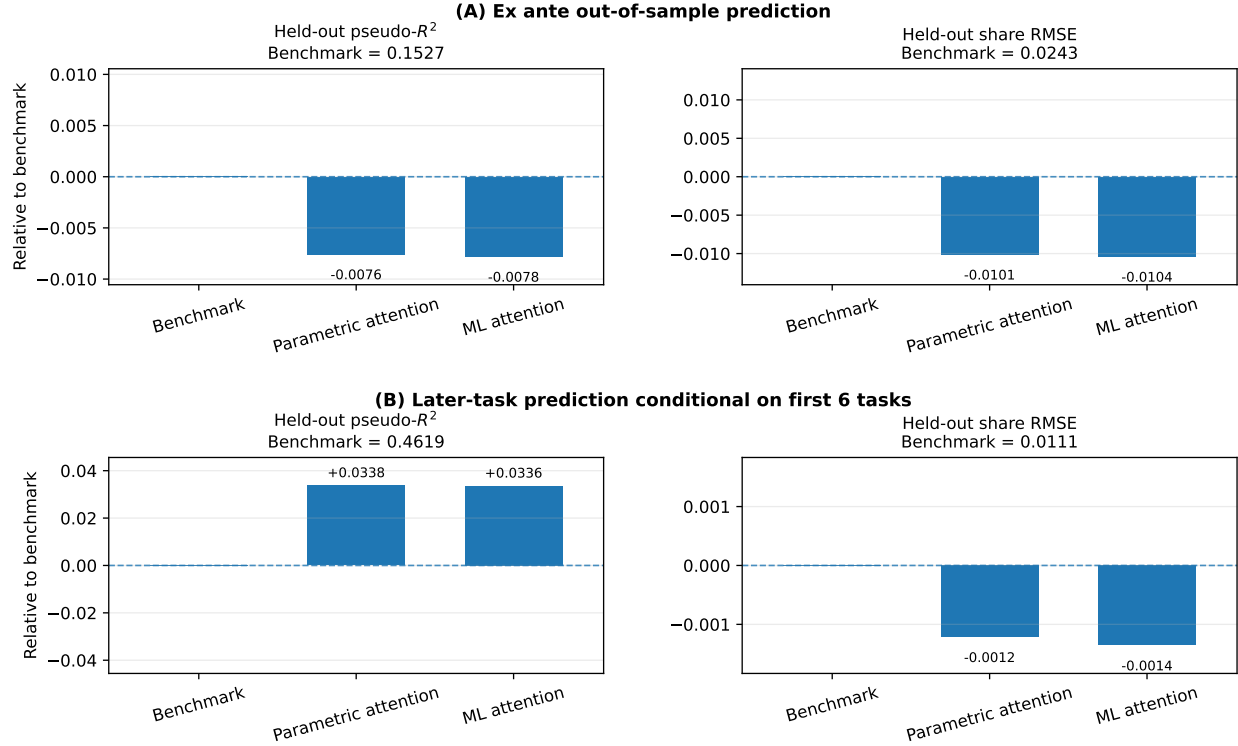


Figure 4: Out-of-sample predictions by model

Notes: Each bar reports the model's performance relative to the benchmark full-attention mixed logit, so positive values indicate improvement over the benchmark and zero denotes the benchmark itself. The left column reports held-out pseudo- R^2 , a respondent-level measure of predictive fit, and the right column reports held-out share RMSE, an aggregate measure of market-share prediction error. Panel (A) uses ex ante out-of-sample prediction for new respondents, while panel (B) predicts later tasks after conditioning on the respondent's first six choices. Because the two metrics are on different scales, comparisons should be made within, not across, columns.

explicitly, the estimated baseline attachment to menthol cigarettes is stronger, substitution patterns are reallocated, and the policy attributes are not assumed to be fully processed by every respondent in every scenario. The net effect is that the full-attention benchmark over predicts cessation and under predicts persistence of menthol demand under prohibition.

For the substantive policy question, if DCEs are used precisely because observed policy variation is unavailable, then the relevant object is the counterfactual. A model that fits the data only by forcing all respondents to attend to all attributes can tell a systematically different story about the likely consequences of regulation.

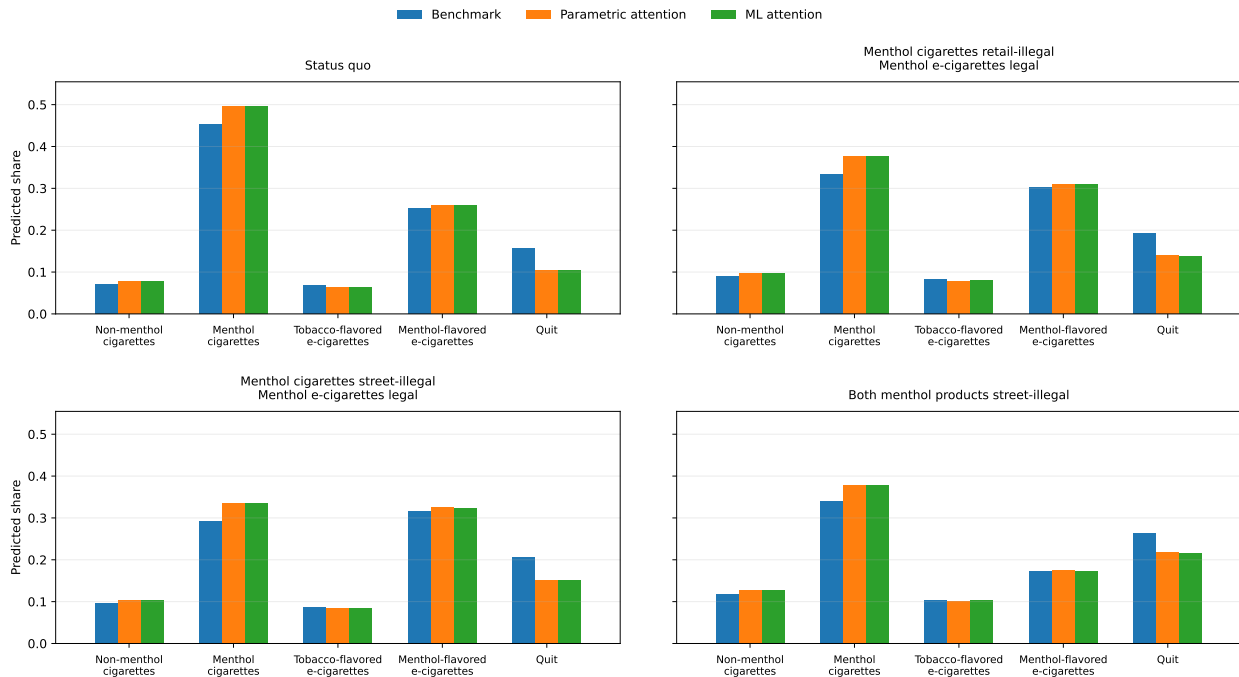


Figure 5: Counterfactual predicted market shares under alternative menthol policy scenarios

Notes: The figure reports model-predicted market shares under four policy scenarios: the status quo; menthol cigarettes retail-illegal with menthol e-cigarettes legal; menthol cigarettes street-illegal with menthol e-cigarettes legal; and both menthol products street-illegal. The scenarios are included to trace how predicted substitution patterns change as the policy environment becomes more restrictive. Shares are averaged over the estimation sample.

6 Discussion and Conclusion

he central methodological lesson is straightforward. In a DCE, muted responses to experimentally varied attributes should not automatically be read as weak preferences or large taste heterogeneity. Some of that attenuation can instead reflect incomplete processing of the attributes used for identification. The menthol DCE application shows that this is not a second-order issue. Once

attention is modeled explicitly, the estimated price and legality coefficients become substantially more negative, conditional on respondents processing those attributes, and the counterfactual implications shift in economically meaningful ways.

At the same time, the validation results argue against an over broad predictive claim. The attention models do not dominate the benchmark in cold-start ex ante prediction. In this application, the strongest empirical case for explicit attention modeling comes from in-sample fit, from the structural interpretation of the coefficients, from conditional prediction after observing respondents' early holdout choices, and above all from the counterfactual policy simulations.

The comparison between the parametric and machine-learning priors is also informative. The machine-learning prior gives additional in-sample fit, but the parametric prior remains highly competitive once model complexity is penalized. That suggests that the economically important regularities in attention are fairly low dimensional in this setting. More flexible priors are still valuable because they discipline robustness, but the evidence does not suggest that a black-box specification is necessary to overturn the substantive conclusions.

Several limitations remain. The recall questions are ex post and noisy, the prior over attention relies on relatively coarse survey-process measures, and the empirical implementation treats attention as respondent-level rather than task-specific because the auxiliary information is measured at the respondent level. In addition, transformed random-coefficient inference in the attention models is better handled by bootstrap than by local Wald approximations alone. None of these limitations overturns the paper's main contribution. Even modest auxiliary information about processing can sharpen structural interpretation and materially alter the policy counterfactuals derived from a DCE.

More broadly, the paper suggests a useful division of labor between standard mixed logit and explicit attention models. A benchmark full-attention model may remain a strong cold-start predictor. But when the research question is structural—how respondents interpret policy attributes, how much of apparent taste heterogeneity is really processing heterogeneity, and what policy counterfactual follows from that decomposition—a measurement-augmented latent-attention model can be worth the additional complexity.

The latent attention state in this paper should be interpreted as attribute-level processing of experimentally varied prices and legality indicators within a displayed choice set. The model does not separately identify alternative-level consideration-set formation. Some inertia patterns could therefore reflect limited consideration or default-specific behavior rather than incomplete processing of attributes alone (Abaluck and Adams-Prassl, 2021; Barseghyan et al., 2021). I defer the comparison between imperfect attention and limited consideration in this context for future research.

References

- Jason Abaluck and Abi Adams-Prassl. What do consumers consider before they choose? identification from asymmetric demand responses. *The Quarterly Journal of Economics*, 136(3):1611–1663, 2021. doi: 10.1093/qje/qjab008.
- Jason Abaluck, Giovanni Compiani, and Fan Zhang. A method to estimate discrete choice models that is robust to consumer search. Technical report, National Bureau of Economic Research Cambridge, MA, 2020.
- Levon Barseghyan, Francesca Molinari, and Matthew Thirkettle. Discrete choice under risk with limited consideration. *American Economic Review*, 111(6):1972–2006, 2021.
- Moshe Ben-Akiva, Daniel McFadden, Kenneth Train, Joan Walker, Chandra Bhat, Michel Bierlaire, Denis Bolduc, Axel Börsch-Supan, David Brownstone, David S. Bunch, Andrew Daly, Andre De Palma, Dinesh Gopinath, Anders Karlstrom, and Marcela A. Munizaga. Hybrid choice models: Progress and challenges. *Marketing Letters*, 13(3):163–175, 2002. doi: 10.1023/A:1020254301302.
- Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. Salience and consumer choice. *Journal of Political Economy*, 121(5):803–843, 2013. doi: 10.1086/673885.
- Andrew Caplin. Measuring and modeling attention. *Annual Review of Economics*, 8(1):379–403, 2016. doi: 10.1146/annurev-economics-080315-015417.
- Andrew Caplin and Mark Dean. Revealed preference, rational inattention, and costly information acquisition. *American Economic Review*, 105(7):2183–2203, 2015. doi: 10.1257/aer.20140117.
- Richard T. Carson and Theodore Groves. Incentive and informational properties of preference questions. *Environmental and Resource Economics*, 37(1):181–210, 2007. doi: 10.1007/s10640-007-9124-5.
- Andrew T. Collins, John M. Rose, and David A. Hensher. Specification issues in a generalised random parameters attribute nonattendance model. *Transportation Research Part B: Methodological*, 56: 234–253, 2013. doi: 10.1016/j.trb.2013.08.001.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38, 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x.

- Verónica Díaz, Ricardo Montoya, and Sebastián Maldonado. Preference estimation under bounded rationality: Identification of attribute non-attendance in stated-choice data using a support vector machines approach. *European Journal of Operational Research*, 304(2):797–812, 2023. doi: 10.1016/j.ejor.2022.04.018.
- Xavier Gabaix. A sparsity-based model of bounded rationality. *The Quarterly Journal of Economics*, 129(4):1661–1710, 2014. doi: 10.1093/qje/qju024.
- Arne Risa Hole. A discrete choice model with endogenous attribute attendance. *Economics Letters*, 110(3):203–205, 2011. doi: 10.1016/j.econlet.2010.11.033.
- Arne Risa Hole, Julie Riise Kolstad, and Dorte Gyrd-Hansen. Inferred vs. stated attribute non-attendance in choice experiments: A study of doctors’ prescription behaviour. *Journal of Economic Behavior & Organization*, 96:21–31, 2013. doi: 10.1016/j.jebo.2013.09.009.
- Don Kenkel, Alan Mathios, Grace Phillips, Revathy Suryanarayana, Hua Wang, and Sen Zeng. Understanding the demand-side of an illegal market: A case study of the prohibition of menthol cigarettes. *Health Economics*, 34(5):956–971, 2025. doi: 10.1002/hec.4937.
- Daniel K. Lew and John C. Whitehead. Attribute non-attendance as an information processing strategy in stated preference choice experiments: Origins, current practices, and future directions. *Marine Resource Economics*, 35(3):285–317, 2020. doi: 10.1086/709440.
- Filip Matějka and Alisdair McKay. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1):272–298, 2015. doi: 10.1257/aer.20130047.
- Geoffrey J McLachlan and Thiriyambakam Krishnan. *The EM algorithm and extensions*. John Wiley & Sons, 2008.
- Riccardo Scarpa, Mara Thiene, and David A. Hensher. Monitoring choice task attribute attendance in nonmarket valuation of multiple park management services: Does it matter? *Land Economics*, 86(4):817–839, 2010. doi: 10.3368/le.86.4.817.
- Dirk Temme, Marcel Paulssen, and Till Dannewald. Incorporating latent variables into discrete choice models — a simultaneous estimation approach using SEM software. *Business Research*, 1(2):220–237, 2008. doi: 10.1007/BF03343535.
- Kenneth E. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge, 2 edition, 2009. doi: 10.1017/CBO9780511805271.

Sander van Cranenburgh, Shenhao Wang, Akshay Vij, Francisco Pereira, and Joan Walker. Choice modelling in the age of machine learning — discussion paper. *Journal of Choice Modelling*, 42: 100340, 2022. doi: 10.1016/j.jocm.2021.100340.

Christian A. Vossler and Sharon B. Watson. Understanding the consequences of consequentiality: Testing the validity of stated preferences in the field. *Journal of Economic Behavior & Organization*, 86:137–147, 2013. doi: 10.1016/j.jebo.2012.12.007.

Appendix

A Estimation Algorithm

A.1 Simulation and parameterization

The benchmark mixed logit uses four normal alternative-specific constants and five negative lognormal coefficients for price and legality. Simulation is based on Halton draws. The specification uses 500 draws for the benchmark model and 200 draws for the latent-attention models. The benchmark is estimated using raw mean and dispersion parameters. Reported means and standard deviations are transformed back to the natural economic scale after estimation.

For each respondent, the coefficient vector is written as

$$\beta_{ir}(\theta) = (\alpha_{ir,1}, \dots, \alpha_{ir,4}, b_{ir,1}, \dots, b_{ir,5})$$

where the four alternative-specific constants are normal random coefficients, $\alpha_{ir,j} = \mu_j + \sigma_j z_{ir,j}$, and the five price/legality coefficients are negative lognormal random coefficients,

$$b_{ir,m} = -\exp(\mu_m + \sigma_m z_{ir,m})$$

This parameterization enforces negative marginal utility for price and illegal-market indicators while still allowing heterogeneity in the magnitude of disutility. The simulated respondent likelihood for the full-attention benchmark is

$$\hat{L}_i^B(\theta) = \frac{1}{R} \sum_{r=1}^R \prod_{t=1}^T \prod_{j=1}^J P_{ijt}(\mathbf{1}_K, \beta_{ir})^{\mathbf{1}_{\{y_{it}=j\}}}$$

The natural-scale means and standard deviations reported in the paper are computed after applying the normal or negative-lognormal transformations, not from the raw optimizer parameters.

A.2 Latent attention state and choice likelihood

The latent state is $s = (s_1, \dots, s_6) \in \{0, 1\}^6$, with one component for each attention dimension: menthol cigarette legality, menthol e-cigarette legality, non-menthol cigarette price, menthol cigarette price, tobacco-flavored e-cigarette price, and menthol-flavored e-cigarette price. For a given state s and coefficient draw r , the systematic utility used in the choice likelihood is

$$V_{ijt}(s, \beta_{ir}) = \alpha_{ir,j} + c'_{ijt} \gamma_{ir} + (s \odot z_{ijt})' \delta_{ir}$$

The corresponding panel choice likelihood is

$$P_{\theta}(y_i|s, \beta_{ir}) = \prod_{t=1}^T \prod_{j=1}^J \left[\frac{\exp\{V_{ijt}(s, \beta_{ir})\}}{\sum_{l=1}^J \exp\{V_{ilt}(s, \beta_{ir})\}} \right]^{\mathbf{1}\{y_{it}=j\}}$$

The respondent-level attention assumption means that s is held fixed across 12 tasks for a respondent. This assumption matches the recall data, which measure whether respondents remembered attribute variation over the full experiment rather than in a particular task.

A.3 Attention prior

The parametric prior can be written as a mixture between a full-attention mass and a product Bernoulli distribution:

$$p_{\psi}(s|W_i) = \lambda \mathbf{1}\{s = \mathbf{1}_6\} + (1 - \lambda) \prod_{k=1}^6 \pi_{ik}^{s_k} (1 - \pi_{ik})^{1-s_k}$$

where $\lambda \in [0, 1]$ is the additional mass on the full-attention state. For the product component,

$$\text{logit}(\pi_{ik}) = a_k + W_i' b_{g(k)}$$

where a_k is a dimension-specific intercept and $g(k)$ maps each dimension to either the price-attention group or the legality-attention group. This structure keeps the prior parsimonious while allowing the three survey-process measures to shift attention differently for price and legality dimensions.

The machine-learning prior replaces the linear index with a one-hidden-layer network. Let $h_i = \phi(B_1 W_i + d_1)$ be the hidden layer, where $\phi(\cdot)$ is a nonlinear activation. Dimension-specific attention probabilities are

$$\pi_i = \sigma(B_2 h_i + d_2)$$

with the logistic transformation $\sigma(\cdot)$ applied componentwise. The prior likelihood is otherwise the same product over attention dimensions, optionally with the same full-attention mass. Regularization is applied only to the prior block. Utility coefficients and measurement probabilities remain structural model objects rather than black-box prediction targets.

A.4 Measurement model

The raw recall responses distinguish four categories: correct recall, incorrect “same” recall, both, and neither. The model uses a ternary coding that combines “same” and “both” into an ambiguous

middle category and retains “neither” as a separate outcome. This choice reduces noise in the measurement block while preserving the basic informational content of the recall items.

Let $R_{ik} \in \{0, 1, 2\}$ denote no recall, ambiguous recall, and correct recall, respectively. For each attention dimension k and latent attention value $a \in \{0, 1\}$, define

$$m_{kac} = P(R_{ik} = c | A_{ik} = a), \quad c \in \{0, 1, 2\}$$

The measurement likelihood is

$$p_\eta(R_i | s) = \prod_{k=1}^6 m_{k, s_k, R_{ik}}$$

Monotonicity is imposed by requiring the measurement distribution under attention to place weakly more mass on responses that are more consistent with attention. In the ternary case, a convenient form is first-order stochastic dominance:

$$P(R_{ik} \geq c | A_{ik} = 1) \geq P(R_{ik} \geq c | A_{ik} = 0) + \epsilon, \quad c = 1, 2$$

where the implementation uses $\epsilon = 0.02$. This restriction fixes the label of the latent state and prevents the measurement block from interpreting correct recall as evidence of inattention.

The tuning choices are as follows. The measurement weight is set to $\omega = 0.1$. The monotonicity gap is set to $\epsilon = 0.02$. The parametric prior uses six dimension-specific intercepts, one shared slope vector for legality attention, one shared slope vector for price attention, and a full-attention mass parameter. The main latent-attention fit uses six EM steps, up to 15 utility iterations within each EM step, and up to 60 prior iterations. Five-fold cross-validation uses a lighter configuration, with 50 draws and shorter numerical updates, to keep the exercise computationally feasible.

A.5 Generalized Expectation-Maximization (EM) algorithm

Let s index the 64 respondent-level attention states and r index the simulation draws used for random coefficients. At EM iteration q , the E-step computes posterior responsibilities

$$\tau_{isr}^{(q)} = \frac{p(s | W_i; \psi^{(q)}) p(R_i | s; \eta^{(q)})^\omega p(y_i | s, \beta_{ir}; \theta^{(q)})}{\sum_{s'} \sum_{r'} p(s' | W_i; \psi^{(q)}) p(R_i | s'; \eta^{(q)})^\omega p(y_i | s', \beta_{ir'}; \theta^{(q)})}$$

Holding $\tau^{(q)}$ fixed, the utility M-step updates θ by maximizing the weighted simulated choice likelihood,

$$Q_\theta = \sum_i \sum_s \sum_r \tau_{isr}^{(q)} \log p(y_i | s, \beta_{ir}; \theta)$$

The measurement M-step updates η by normalized posterior counts, followed by the monotonicity projection. The prior M-step updates ψ either by maximizing the shared-slope logistic prior (parametric model) or by maximizing a penalized neural network objective. Because the utility and prior updates are solved numerically rather than analytically, the procedure is a generalized EM algorithm.

For the measurement block, let $\bar{\tau}_{is} = \sum_r \tau_{isr}$ and define smoothed posterior counts

$$N_{kac} = \alpha_c + \sum_i \sum_s \bar{\tau}_{is} \mathbf{1}\{s_k = a\} \mathbf{1}\{R_{ik} = c\}$$

where α_c is a small Dirichlet smoothing constant. Without monotonicity, the closed-form update is

$$\hat{m}_{kac} = \frac{N_{kac}}{\sum_{c'=0}^2 N_{kac'}}$$

The final update projects these probabilities onto the monotonicity-constrained set. This step is equivalent to an isotonic adjustment of the cumulative probabilities and is applied separately by attention dimension.

For the prior block, the posterior weight on state s is $\bar{\tau}_{is} = \sum_r \tau_{isr}$. The parametric update solves

$$Q_\psi(\psi) = \sum_i \sum_s \bar{\tau}_{is} \log p_\psi(s|W_i)$$

while the flexible prior solves

$$Q_\psi^{ML}(\psi) = \sum_i \sum_s \bar{\tau}_{is} \log p_\psi(s|W_i) - \lambda_{ML} \mathcal{P}(\psi)$$

where $\mathcal{P}(\psi)$ is the network penalty. These updates use the EM posterior as soft labels for the latent attention states.

A.6 Numerical stabilization and validation

All posterior calculations are implemented on the log scale using the log-sum-exp normalization. This is important because the panel likelihood multiplies probabilities across 12 choice tasks and can become numerically small. Within a given estimation run, simulation draws are held fixed across EM iterations so that changes in the objective reflect parameter updates rather than simulation noise.

The validation exercises mirror the likelihood decomposition. In the 5-fold ex ante exercise, held-out respondents are predicted using $p_\psi(s|W_i)$ and the estimated mixing distribution for β_i , their choices and recall measures do not update the posterior before prediction. In the adaptive later-task exercise, the posterior is updated using the respondent's first six choices, and the resulting

posterior predictive distribution is used for tasks 7-12. This distribution explains why the attention models can underperform the benchmark in cold-start prediction while improving probability-based prediction after within-respondent learning.