Political Analysis (2025), **00**, 1–14
doi:10.1017/pan.2024.34

PA

ARTICLE

Crowdsourced Adaptive Surveys

Yamil Ricardo Velez

Department of Political Science, Columbia University, New York, NY, USA; Email: yrv2004@columbia.edu

(Received 16 May 2024; revised 21 August 2024; accepted 12 December 2024)

Abstract

Public opinion surveys are vital for informing democratic decision-making, but responding to rapidly changing information environments and measuring beliefs within hard-to-reach communities can be challenging for traditional survey methods. This paper introduces a crowdsourced adaptive survey methodology (CSAS) that unites advances in natural language processing and adaptive algorithms to produce surveys that evolve with participant input. The CSAS method converts open-ended text provided by participants into survey items and applies a multi-armed bandit algorithm to determine which questions should be prioritized in the survey. The method's adaptive nature allows new survey questions to be explored and imposes minimal costs in survey length. Applications in the domains of misinformation, issue salience, and local politics showcase CSAS's ability to identify topics that might otherwise escape the notice of survey researchers. I conclude by highlighting CSAS's potential to bridge conceptual gaps between researchers and participants in survey research.

Key words: large language models; natural language processing; multi-armed bandits; misinformation; public Opinion

Edited by: Jeff Gill

Survey research plays a critical role in informing political decision-making (Page and Shapiro 1983). However, surveys can be too slow to adapt to changing information environments, especially in societies marked by social and political heterogeneity. First, surveys may struggle to identify and prioritize emerging issues. This is especially salient in the context of elections, where campaign-related events, foreign policy incidents, or unexpected economic developments are typical occurrences. The inherent lag time between recognizing issues as they arise and fielding a survey can lead to missed opportunities for capturing real-time opinion shifts. Second, there is often uncertainty about which questions from a larger set should be included in a survey. Third, there can be a disconnect between how researchers and participants interpret the same survey questions - an issue that is especially acute when studying minority groups, whose perspectives may diverge from mainstream viewpoints.[1]

In this paper, I propose a crowdsourced adaptive survey methodology (CSAS) that leverages advances in natural language processing and adaptive algorithms to create participant-generated questionnaires that evolve over time. I use open-ended responses from participants to create question banks comprised of potential survey questions, from which questions are prioritized using a multi-armed bandit algorithm. While the survey is in the field, participants contribute to the question bank and respond to questions submitted by other participants, enabling the algorithm to explore and prioritize survey questions that resonate with the sample. Even in well-trodden settings such as identifying salient issues, the CSAS method produces promising items that warrant further investigation (see Figure 1 for a summary of the process and representative issues recovered using the method).

[1]For instance, Anoll (2018) shows that minority groups' lived experiences and collective beliefs can lead to different understandings of concepts like political engagement.

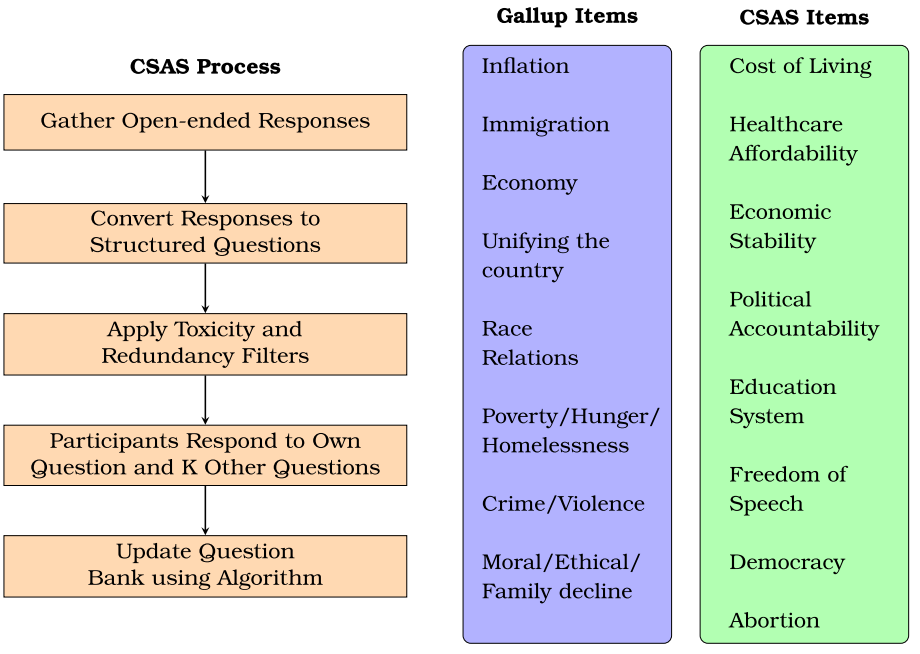

Figure 1. CSAS process flowchart and representative set of issues uncovered by CSAS.

I apply this methodology across three domains: gauging the prevalence of false beliefs among Latinos, measuring issue salience, and identifying local political concerns. First, I use CSAS to develop an evolving battery of issues to gauge issue importance in the aggregate. Despite seeding the algorithm with Gallup issue categories, I find that popular issue topics based on open-ended responses depart from the set of "most important issues," reflecting concerns over healthcare, inflation, political accountability, and crime.[2] Then, I move to more niche applications: misinformation within the Latino community and local political concerns. I find that CSAS reveals claims and issues that would likely escape the notice of survey researchers.

The advantages of CSAS are threefold. First, it enables survey researchers to capture trends in public opinion in real-time, reflecting the public's evolving beliefs and concerns. Second, it allows survey researchers to apply a more inductive approach to questionnaire construction. Finally, it democratizes the survey process by allowing respondents to contribute to instruments. These benefits come at little cost in terms of survey length. Researchers can set the number of "dynamic questions" in advance, and select the appropriate algorithm for determining how questions should be prioritized. For example, one can rely on a set of tried-and-true items, while allocating a few survey slots for dynamic questions.

Beyond applications to misinformation and issue importance, the method can be applied to research topics such as social movements, public reactions to unfolding political events, and political representation, among others. By enabling researchers to adapt survey instruments to changing information environments and democratizing the survey process, the CSAS approach could provide new insights into political behavior and complement traditional survey methods across various sub-fields in political science.

[2]While some Gallup codes (e.g., "ethics, moral, and family decline") are not expressed in everyday language, complicating the comparison with CSAS-generated summaries, many codes (e.g., "immigration," "poverty," "crime") closely align with CSAS outputs.

1. Dynamic Survey Methodologies: Existing Approaches

Influential texts on survey design stress the importance of a "tailored" approach to recruitment and stimuli, where survey materials are adapted to populations under study (Dillman, Smyth, and Christian 2014). Adapting questionnaires to respondents can enhance measurement and participant satisfaction. For instance, branching questions can reduce respondent burden and measurement error by eliminating irrelevant sub-questions (Dillman and Christian 2005; Krosnick and Berent 1993). Questions measuring recognition of elected officials and voting in sub-federal elections can be tailored using location to produce more relevant question content *via* "piped in" text that is automatically shown to participants in online surveys (Ansolabehere and Schaffner 2009). These examples illustrate how surveys already possess dynamic elements that respond to user input or data.

1.1. Computerized Adaptive Tests

Scholars have recently developed methods for carrying out computerized adaptive tests (CATs) in public opinion surveys (Montgomery and Cutler 2013; Montgomery and Rossiter 2020, 2022). CAT algorithms "respond to individuals' prior answers by choosing subsequent questions that will place them on the latent dimension with maximum precision and a minimum number of questions." (p. 173). CATs are typically employed using latent scales, where the goal is to optimize the number of questions. Montgomery and Cutler show that CATs offer a superior approach to traditional static batteries, and these tools can be easily implemented in survey software such as Qualtrics (Montgomery and Rossiter 2022).

CATs rely on pre-existing measurement scales (e.g., political knowledge, personality batteries). However, in settings where the objective is to capture novel issues or changes in the information environment, scholars and practitioners may want to learn about the prevalence of discrete beliefs, some of which cannot be known in advance. Thus, while CATs allow us to enhance precision when estimating latent traits or ideal points, there are settings where the question bank cannot be fixed in advance and describing the nature of discrete items, rather than estimating positions on latent scales, is the primary objective. These settings include misinformation research, where survey questions draw from fact-checking sources and social media data, and studies of elections and campaigns, where events and shifts in news coverage are integral to understanding the dynamics of races.

1.2. Wiki Surveys

Wiki surveys are a collaborative survey methodology where users help shape instruments (Salganik and Levy 2015). Drawing inspiration from online information aggregation portals such as Wikipedia, wiki surveys balance three principles: greediness, collaborativeness, and adaptivity. Greediness refers to capturing as much information as respondents are willing to provide; collaborativeness refers to allowing respondents to modify instruments (e.g., proposing new items); and adaptivity refers to optimizing instruments to select the "most useful information." While wiki surveys have shown promise in facilitating collective decision making (e.g., allowing users to vote on policies—both pre-determined and user-generated—that should be considered by local governments), existing applications rely on pairwise comparisons between options provided by survey designers and participants. However, pairwise comparisons may not be useful in settings where options can be accorded the same weight, the decision is not zero-sum, and outcomes can be more accurately measured on an ordinal or continuous scale.

2. The Crowdsourced Adaptive Survey Method

Building on the wiki survey and other attempts to insert dynamic elements into existing surveys (e.g., CAT approaches), I develop a method that enables question banks to evolve based on user input and

Table 1. Overview of the CSAS methodology.

Step	Model	Details
Collect open-ended responses	-	Participants provide responses to open-ended questions.
Process open-ended responses	LLM (e.g., OpenAI's GPT-3.5)	Open-ended responses are processed and converted into structured survey questions using LLMs.
Apply filters	Document embeddings (e.g., OpenAI's *text-embedding-ada-002*)	*If flagged as redundant,* do not add to question bank.
	Toxicity detection (e.g., OpenAI's moderation endpoint)	*If flagged as toxic,* do not add to question bank.
		If not flagged as redundant or toxic, add structured text to question bank.
Participant ratings	-	Participants rate their own question and *k* other questions selected using a multi-armed bandit.
Update question bank	Multi-armed Bandit algorithm (e.g., Gaussian Thompson sampling)	The question bank is updated using ratings, prioritizing the most highly rated questions.

does not impose constraints on question formats. I use generative pre-trained transformers (GPTs) to convert participants' open-ended responses into structured questionnaire items (see Velez and Liu (2024) for an example). I then employ adaptive algorithms (Offer-Westort, Coppock, and Green 2021) to identify the "best-performing" questions from the question bank.[3] First, each respondent answers an open-ended question about a given topic that is cleaned, summarized, and converted into a structured survey question format. Second, respondents respond to their own questions, along with *k* other questions from a question bank generated by previous participants. Finally, ratings for user-submitted questions and *k* questions drawn from the question bank are updated using a multi-armed bandit algorithm, adjusting the probabilities of presenting these questions to future participants in subsequent surveys.[4]

The three essential features of the proposed method are open-ended questions, a question bank, and an algorithm for updating question selection probabilities. Table 1 displays the different steps of the CSAS method: eliciting potential items using open-ended questions; processing and filtering candidate items; and optimizing question selection. I walk through each step in turn.

The open-ended question is used to query participants in a free-form manner about a given topic, issue, or claim. Since these data will be unstructured, introducing heterogeneity on dimensions such as length, style, and grammar, a response conversion stage is typically necessary. For example, given that Likert scale questions in public opinion surveys tend to be brief, one can convert an open-ended response of arbitrary length into a sentence-long summary using a GPT. GPTs are large-parameter language models that can perform various tasks such as text prediction, summarization, and classification at levels that mirror human performance (Radford *et al.* 2018; Vaswani *et al.* 2017). Recent years have seen the development of a diverse range of GPTs, with proprietary models like OpenAI's GPT-4, Anthropic's Claude, and Google's Gemini demonstrating robust capabilities across

[3]Questions are deemed "best-performing" based on their ability to maximize researcher-defined objectives (e.g., mean accuracy estimates, mean importance ratings).

[4]Following the "greediness" principle of wiki surveys, users rate their own questions to maximize the information collected about items.

various text manipulation, summarization, and generation tasks. In parallel, "open source" and "open weights" models, including Meta's LLaMA and Mistral AI's Mixtral, have emerged, offering competitive performance in similar domains (see Appendix E of the Supplementary Material, for a detailed discussion and direct comparison of these models).

Once open-ended response data are in a usable, structured format, they can be included in a question bank. Though inclusion of questions can be unrestricted at this stage, researchers may want to impose additional constraints to reduce redundant questions and apply filters to ensure that survey questions meet the researcher's objectives (e.g., reducing toxicity, increasing relevance).

Focusing on redundancy first, if two respondents submit responses about Democratic spending priorities with only minor differences in wording, it may be unnecessary to include both questions. Moreover, multi-armed bandit algorithms can struggle to identify the best-performing arm when arms are equally matched (i.e., a "no clear winner" scenario). Assuming near-identical questions are rated similarly, this increases the odds of failing to identify the best-performing item (Offer-Westort *et al.* 2021, 832–833). Because open-ended responses are rarely exact duplicates, more sophisticated NLP methods (e.g., document embeddings) can be used to filter near-duplicates. Document embeddings locate texts on a high-dimensional space and can be used to identify similarities between texts (Rheault and Cochrane 2020). By applying document embeddings, researchers can quantify the similarity between different questions even when the wording varies, and retain only questions that surpass a pre-defined threshold of uniqueness.

Researchers may also want to apply additional filters to ensure that questions meet pre-specified criteria on dimensions such as relevance and toxicity. For example, for a survey of issue importance, a survey researcher may choose to exclude responses referring to the personal characteristics of politicians. Given that this is a classification task, one may opt for a supervised learning model trained on a labeled dataset or a GPT model, among other options. The same holds for identifying and removing toxic responses. Given that a small percentage of respondents resort to "trolling" behavior, ensuring other participants are not exposed to harmful content is important (Lopez and Hillygus 2018).

The next challenge lies in how questions are presented and selected within the survey, balancing between identifying "best-performing items" (exploitation) and examining the full set of candidate items (exploration). This entails the use of an algorithm that takes a set of candidate items and determines how those items are presented to participants. In some domains, such as misinformation or issue salience, prioritizing popular claims or important issues may be crucial. Conversely, in fields like personality research, capturing the full spectrum of trait variation, including low-prevalence items, is more important. In the latter case, uniform sampling from the question bank could be useful (see Appendix J of the Supplementary Material).

Across the three applications presented in this paper, I focus on identifying the "best-performing" survey questions, defined as those with higher mean scores for the respective measures. I employ multi-armed bandit (MAB) algorithms, specifically Gaussian Thompson sampling, to prioritize item selection. These algorithms balance two key objectives: exploitation (focusing on questions that show promise) and exploration (testing additional questions to assess their potential). MAB algorithms typically use outcome measures to guide the assignment of participants to different conditions (or, in our case, questions), prioritizing those that score higher on predetermined metrics (e.g., mean scores). Through this iterative process, Thompson sampling efficiently allocates more participants to questions that "resonate" more with the sample.

3. An Application to Issue Salience

In contrast to conventional approaches that use pre-defined issues to study issue importance (see Ryan and Ehlinger (2023) for a critique), CSAS can be used to produce a dynamic slate of issues. This can be helpful in estimating support for idiosyncratic issues that may not appear on the national agenda, but still inspire strong reactions among "issue publics" (Elliott 2020) or serve as issues that could be mobilized in the future, corresponding to the elusive notion of "latent opinion" described in Key (1961).

Since 1935, Gallup's Most Important Problem (MIP) question has been used to identify the issue priorities of the American public. Using an open-ended format, participants are asked "What do you think is the most important problem facing this country today?," with responses being hand-classified into a set of categories corresponding to broad issue or policy areas. Despite its adoption in public opinion research, the measure has been criticized for being an imperfect proxy of salience. As Wlezien (2005) argues, the question asks respondents to provide information on two distinct concepts—importance and "problem status." While some respondents may interpret the question as one where they can offer a personally relevant issue, others may interpret it as an opportunity to highlight a problem affecting the nation as a whole.

More recently, Ryan and Ehlinger (2023) make a case for moving beyond both fixed closed-ended questions and hand-coded classifications of open-ended questions when measuring issue importance. Like the MIP, Ryan and Ehlinger (2023) use open-ended questions to elicit issue positions from participants. However, they ask participants to directly reflect on issues of personal importance. Moreover, in contrast to the typical MIP method, they obtain closed-ended measures of personal importance for the issues elicited using the open-ended method. This approach provides a richer amount of information about the *degree* to which an elicited issue may hold importance to an individual. Recent studies in this vein have recovered high levels of stability (Velez and Liu 2024) and sizable causal effects for this "core issue" in conjoint settings (Ryan and Ehlinger 2023; Velez 2023).

Applying CSAS in this context is straightforward. Participants report personally relevant issues using open-ended responses[5]; open-ended responses are transformed into a structured question and included in a question bank; and participants respond to the different issue topics in an issue importance battery, with Gaussian Thompson sampling used to optimize question selection. I seed the question bank with a set of eight Gallup items that were popular at the time and assess whether the crowdsourced issue topics receive higher importance ratings. The issues were the following: "Immigration", "Economy", "Race Relations", "Poverty", "Crime", "Ethics, Moral, and Family Decline", "Unifying the country", and "Inflation."

From September 11 to September 13, 2023, I collected data from a national quota sample balanced on age, race, and gender using CloudResearch Connect (N = 820). CloudResearch Connect provides a low-cost, high-quality online convenience sample. Its demographic representativeness is comparable to other popular sample providers such as Prolific (Stagnaro *et al.* 2024).[6] OpenAI's GPT-4 was instructed to convert unstructured text into issue topics and filter out irrelevant or redundant topics.[7] This was accomplished using a retrieval-augmented generation (RAG) pipeline. This process involved retrieving the five nearest neighbors for each potential issue, allowing GPT-4 to consider similarity and avoid redundancy when generating new issue items. For each submission, I used OpenAI's *ada* embeddings model to generate a 1536-dimensional embeddings vector and retrieved the five nearest neighbors using a vector database.[8] GPT-4 was instructed to avoid generating issues similar to the "nearest neighbors" that were retrieved (see Appendix D of the Supplementary Material).

Gaussian Thompson Sampling (GTS) was then used to determine which questions to present to subsequent participants. Though traditional Thompson Sampling requires binary outcomes due to its use of the Beta distribution (Offer-Westort *et al.* 2021), GTS can be leveraged to accommodate continuous outcomes (Agrawal and Goyal 2017).[9] GTS was implemented in real-time using a custom-built back-end system created by the author. In contrast to previous applications of adaptive experiments

[5]The wording of the open-ended question is based on the open-ended question described in Ryan and Ehlinger (2023).

[6]CloudResearch Connect participants score at the upper end of attentiveness, which may not reflect all survey environments. For samples with lower attentiveness, researchers could implement additional measures such as conditioning inclusion of participant-generated questions on attention check performance or using open-ended response quality checks.

[7]When rating their own issues, the issue "room temperature superconductors" was presented to those submitting gibberish or nonsense responses.

[8]I used Pinecone's API, a vector database service, to store and retrieve embeddings.

[9]As in other research (Offer-Westort *et al.* 2021), a probability floor of .01 was employed to guarantee that every item in the item bank retained a non-zero and non-negligible chance of being presented to participants.

in political science that have leveraged batch-wise Thompson sampling, probabilities were adjusted at the respondent level.

Recent studies have shown that the efficacy of GPTs such as GPT-4 in generating issue categories using open-ended text is on par with classification algorithms trained on thousands of examples, achieving performance levels marginally below that of human evaluators (Mellon *et al.* 2024). Each participant rated their own issue, along with eight others that were determined using GTS. Issue importance was measured using the question, "How important is this issue to you personally?" Responses were recorded on a five-point ordinal scale with the following options: "Not at all important," "Slightly important," "Moderately important," "Very important," and "Extremely important."

3.1. Results

Figure 2 displays mean estimates for issue topics receiving 50 or more issue importance ratings.[10] As shown in the figure, the highest-rated issues were focused on the economy and health care, with issues such as "Cost of Living," "Healthcare Affordability," "Healthcare Costs," "Economic Stability," and "Universal Healthcare." Issues rated lower on importance include more social and culture issues such as immigration and voting rights.[11] The MIP issue topics of "Race Relations" and "Ethics, Moral, and Family Decline" appeared among the lower-rated items, as did topics related to immigration (i.e., "Illegal Immigration," "Border Security"). In the list of highly rated topics, we see issues that would likely not appear in traditional issue importance batteries such as "Mental Health Access," "Privacy Protections," and "Candidate Transparency." Moreover, the frequent mention of various economic and healthcare dimensions is instructive, indicating the salience of economic concerns in the sample. In Appendix B of the Supplementary Material, I assess heterogeneity across partisan subgroups finding some issues that generate consensus across parties such as healthcare costs, political polarization, and economic issues.

4. An Application to Latino Information Environments

Moving from issue importance to concerns within hard-to-reach communities, I use the CSAS method to identify rumors, negative political claims, and misinformation reaching Latinos, a group that has received attention among journalists and social scientists due to potential misinformation campaigns targeting the community (Cortina and Rottinghaus 2022; Velez, Porter, and Wood 2023). I focus on Latinos for two reasons. First, fact-checking is still a relatively new institution within Latino-oriented media (Velez *et al.* 2023, 790). Existing organizations might overlook important claims that circulate within the community due to resource constraints and the possibility that best practices for verification are still being developed. Second, private, encrypted messaging applications used by Latinos such as WhatsApp and Telegram may hinder the detection of false claims. In contrast to misinformation that is transmitted through social media such as Instagram, Twitter, and Facebook, fact-checkers and researchers may not be privy to claims circulating in private channels.

Implementing this more "bottom-up" approach to misinformation detection, I fielded a survey of 319 self-identified Latinos from the United States using the survey platform, CloudResearch Connect, from July 6 to 7, 2023. First, participants were asked two open-ended question regarding negative claims they had heard about Republicans and Democrats.[12] These claims were then passed to a fine-tuned

[10] Mean estimates are weighted by the inverse probability of selection, as in Offer-Westort *et al.* (2021).

[11] Though healthcare-relevant issues could be collapsed into a single item, allowing for more detailed issue categories enables researchers to explore different dimensions of the same issue that may be important to participants. Moreover, if more coarse issue categories are desirable, these can be incorporated into the LLM's prompt using few-shot prompting (e.g., including multiple examples in the prompt that produce the intended classifications).

[12] This approach expands the inclusion criteria for the question bank beyond the traditional definition of misinformation to include falsifiable statements that portray parties or candidates negatively. While unconventional, this broader inclusion criteria allows us to cast a wider net, revealing interesting distinctions between newsworthy scandals and false claims that might otherwise go unnoticed. Future iterations could employ more sophisticated methods, such as comparing open-ended

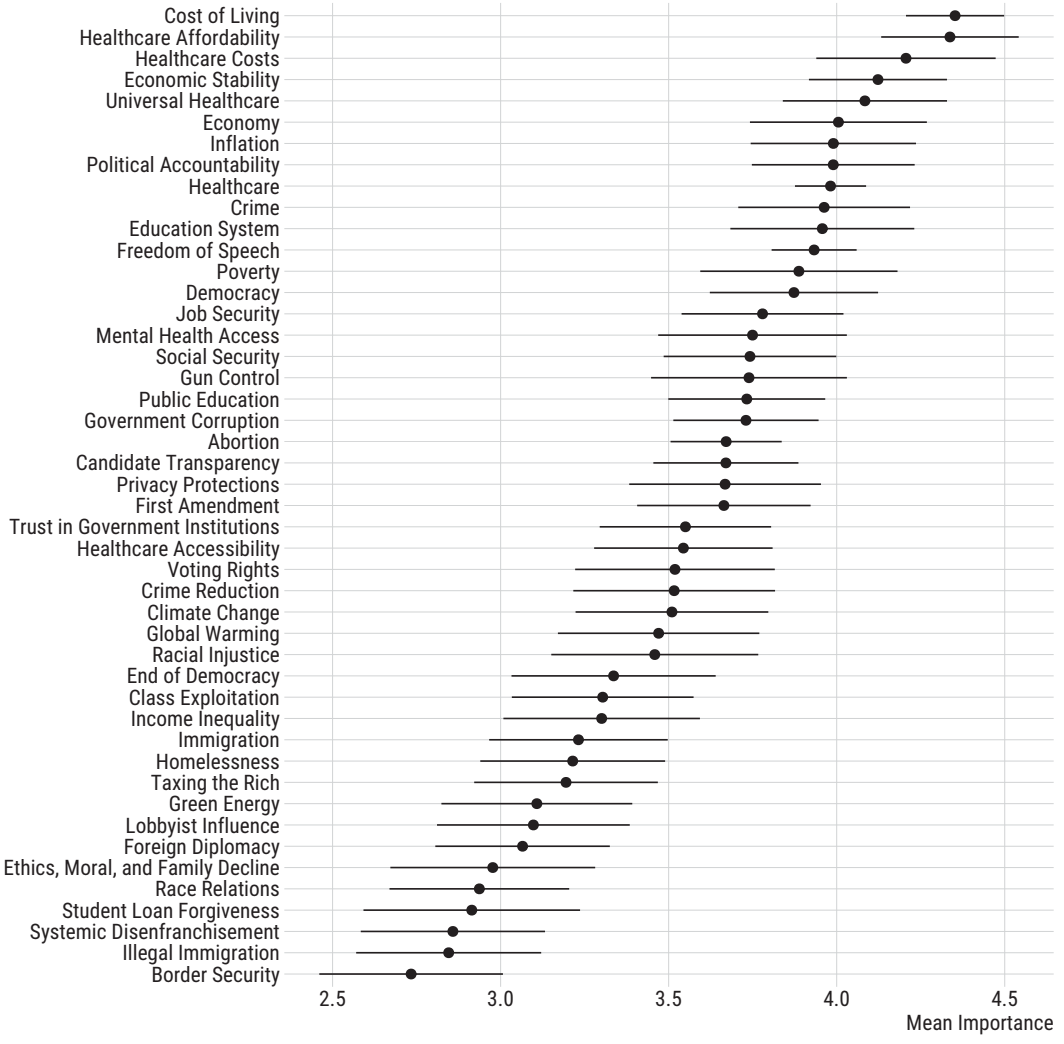

Figure 2. IPW-weighted estimates of survey questions measuring issue importance with corresponding 95% confidence intervals.

version of OpenAI's *ada* text generation model that classified the text as a "verifiable claim."[13] Fine-tuning was necessary to ensure that questions entering the question bank were falsifiable political claims, rather than value judgments (e.g., politicians are evil). To carry out the fine-tuning step, a mixture of researcher-provided examples and participant-provided examples (N = 87) were hand-coded to indicate whether a claim was falsifiable in principle. Hand-coded classifications were then used to fine-tune the *ada* text generation model.[14]

responses against databases of fact checks to isolate verifiably false statements. Additionally, one could explore alternative objectives, like identifying items that are factually true but commonly misperceived as false.

[13]Note that there is a distinction between the ada embeddings and text generation models.

[14]Since this study was conducted, more advanced models have emerged. These models may obviate the need for fine-tuning through the use of few-shot prompting (Chen, Yi, and Varró 2023), where a selection of examples can be included in the prompt by the researcher to guide the model's behavior.

I also used a similarity and toxicity filter before adding items to the question bank.[15] For each submitted question, I used the *ada* model in OpenAI's embeddings API, retrieved the five nearest neighbors using a vector database, and filtered out questions with a similarity score above .90.[16] I also used OpenAI's moderation endpoint to filter out offensive and toxic claims. Claims that passed these filtering steps were added to the question bank.

After the submission and cleaning step, participants responded to their own question bank submission, along with four items from the question bank and six items capturing conspiracy beliefs and more common misinformation items (e.g., Covid-19 vaccines modify your DNA). All of the questions were presented in a matrix format, with a four-point accuracy scale. The response options were "not at all accurate," "not very accurate," "somewhat accurate," and "very accurate." Four claims were taken from the front pages of Latino-oriented fact-checking websites (i.e., Telemundo's T-Verifica, Univision's El Detector) to seed the question bank. As in the issue importance study, GTS was used to update question selection probabilities with accuracy ratings being used as the metric for updates.

5. Results

Figure 3 displays mean accuracy estimates for claims receiving more than ten ratings. The claims with the highest perceived factual accuracy covered information about both Republicans and Democrats. Party stereotypes about Republican positions on moral issues were rated as highly accurate ($\bar{x}$ = 3.54; SE = .07), as were claims that Trump used the presidency to enrich family and friends ($\bar{x}$ = 3.54; SE = .16) and possessed classified documents in his vacation homes ($\bar{x}$ = 3.43; SE = .12). Other highly rated claims focused on Republicans such as Ted Cruz ($\bar{x}$ = 3.40; SE = .05), extreme policy positions such as a claim that Trump plans to deport all undocumented people ($\bar{x}$ = 3.32; SE = .31), and President Biden's son, Hunter Biden ($\bar{x}$ = 3.24; SE = .20). The lowest-rated claims typically involved false statements, allegations, or extreme descriptions of issue positions. Claims that scored especially low on perceived accuracy included "This year, there was a major explosion close to the White House" ($\bar{x}$ = 1.50; SE = .15), "There were no wars during the Trump presidency" ($\bar{x}$ = 1.74; SE = .17), "Republicans support going back to the 1776 constitution." ($\bar{x}$ = 1.80; SE = .33), "Trump will deport **all** immigrants if he wins in 2024" ($\bar{x}$ = 1.91; SE = .19), and "The president was involved with his son's bribes" ($\bar{x}$ = 1.92; SE = .20). It is worth noting that the standard errors increase as mean factual accuracy ratings *decrease*, and vice versa. This stems directly from the Gaussian Thompson sampling algorithm, which directs more participants to rate 'promising' items.

Whereas claims rated highly on perceived accuracy mostly reflected actual events or generalizations of party positions, claims with lower accuracy ratings typically involved verifiably false information or oversimplifications. These findings are instructive in that they reveal a level of discernment in the aggregate. Objectively false claims are generally seen as less credible by participants. Instead, higher accuracy ratings are observed when participants judge claims that are widely reported in the news (e.g., Ted Cruz's Cancun trip amid a Texas blackout) or that reflect commonly-held perceptions of party positions (e.g., Republicans opposing abortion and same sex marriage). Though the initial seed claims based on fact-checks were small in number, the analysis revealed a surprising disparity: the most readily believed claims were not identified by the fact-checking organizations, but rather originated from other participants. In Appendix B of the Supplementary Material, I explore heterogeneity by levels of trust in the private, encrypted messaging application WhatsApp, which has been argued to be a vector of misinformation (Velez *et al.* 2023). I find that those who trust the platform are less likely to view genuine

[15] If a user submission did not successfully pass these filters, respondents did not rate their own submission, but instead were asked to rate generic items about Republicans (Democrats) being too conservative (liberal).

[16] In initial tests before data collection, lower similarity thresholds such as .80 were found to produce false positives (e.g., classifying "Biden is a tax cheat" and "Trump is a tax cheat" as sufficiently similar) and higher similarity thresholds such as .95 produced false negatives (see Appendix H of the Supplementary Material, for a discussion of different methods for reducing redundant items).

https://doi.org/10.1017/pan.2024.34 Published online by Cambridge University Press

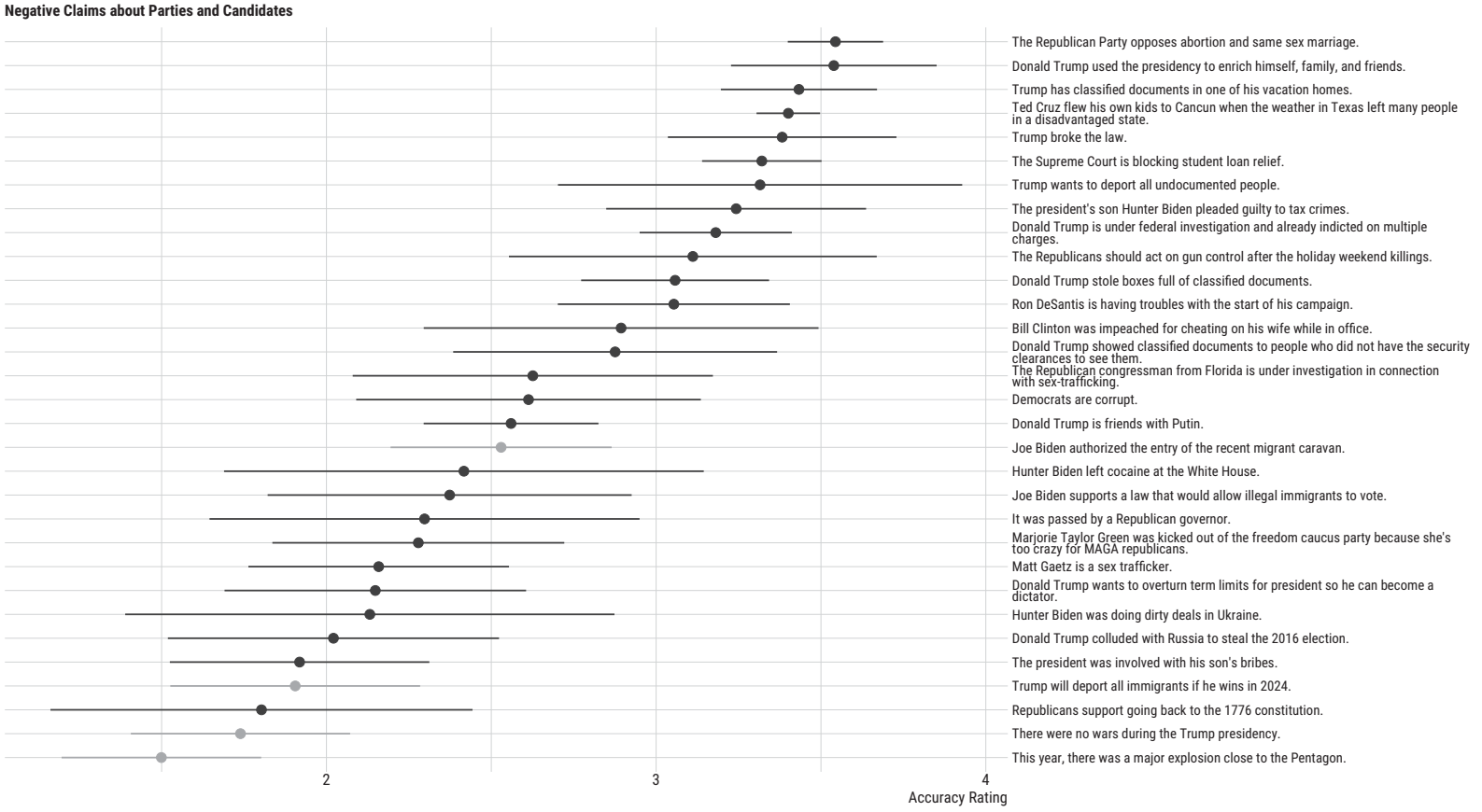

Figure 3. IPW-weighted estimates of survey questions measuring negative political claims with corresponding 95% confidence intervals. Items in gray were initial seed items based on fact-checked claims produced by Latino-focused fact-checking organizations. Black items are participant-generated items.

newsworthy scandals as accurate, but rate factually inaccurate claims similarly to those who do not trust WhatsApp. This suggests that gaps between the two groups might not be a function of misinformation *per se*, but instead more general political knowledge.

The results highlight the complexity of the information environment among Latinos in the United States. They engage with a variety of narratives, some of which portray different parties negatively, but also reflect actual events. These findings offer a glimpse into how a prominent and politically pivotal ethnic group, which receives significant attention from campaigns, interfaces with information about political parties. Moving beyond the application to Latinos, an advantage of CSAS is that group-specific understandings may be reflected in the question bank and researchers can assess whether questions that perform well within certain subgroups also perform well in others. Moreover, while the focal outcome of this study is perceived accuracy, one could implement variations of CSAS for misinformation measurement that optimize for exposure. In this setup, participants might be asked to report whether they have seen a set of claims. Though exposure does not equate to belief, this approach could help identify false claims that are gaining traction within specific communities.

6. An Application to Local Political Issues

In Appendix K of the Supplementary Material, I examine local political concerns. Similar to the case of studying minority opinion, inconsistent polling at the local and state levels can complicate the development of quality survey items. Dovetailing with recent research on the "nationalization" of local politics (Hopkins 2018), policy domains that are typically considered national in scope such as immigration, gun policy, foreign policy (e.g., stances on Gaza), and the environment emerge as participant-submitted items and gain traction within the sample. In surveys of particular states, Congressional districts, or cities, CSAS could be a valuable tool for uncovering local attitudes, beliefs, and preferences.

7. Concerns and Caveats

7.1. Is the CSAS method compatible with traditional survey design?

Despite the limitations of traditional surveys in identifying changing information environments or measuring responses within minority or otherwise hard-to-reach communities, the two approaches are not at odds. Researchers can decide the number of adaptive questions, and include these questions in standard batteries. For example, in the Latino survey, participants rated a pre-existing set of false claims and conspiracies, along with an adaptive set, in a question matrix. Before using this method, scholars should determine whether the marginal benefit of having a designated slot for exploratory questions is worth the survey time and cost.[17] A distinct advantage of multi-armed bandits is that several items can be explored despite having a smaller set of survey slots. The two approaches can also work in tandem when there are multiple phases of data collection. An initial wave (or pilot) could use CSAS to develop a fixed battery of questions for future waves, functioning much like pilot studies that gather open-ended data to inform scale construction (Weller 1998). With CSAS, however, future surveys can be designed not only with open-ended content in hand, but question ratings and posterior probabilities that a given question is the best-performing question. This approach may be optimal if researchers prefer to split their research process into exploratory and confirmatory stages, as is recommended in Egami *et al.* (2018). Pre-registration across these different stages could lend more credibility to conclusions derived using this method (see Offer-Westort, Rosenzweig, and Athey (2022) for an example).

7.2. Late Arrivals, Prompting Participants, and Costs of LLM Inference

In Appendix C of the Supplementary Material, I address additional issues such as "late arrivals" or items added to the question bank toward the end of the survey, general concerns about crafting open-ended

[17] In the 2024 ANES pilot, the median response time for an open-ended issue importance question was 17 seconds.

questions, and inference costs. I estimate inference costs for both closed-source and open-source models (i.e., OpenAI's GPT-4 versus Mistral AI's Mixtral), finding that costs ranged between $0.005 and $0.01 per participant. Furthermore, I discuss the issue of toxicity and recommend using moderation APIs. In Appendix G of the Supplementary Material, I provide detailed steps for implementing the CSAS method. I develop a Django application that can be hosted on popular platforms such as Amazon Web Services, Replit, or Google Cloud. Finally, in Appendix I of the Supplementary Material, I examine whether sample composition varies over the course of the study but fail to detect evidence of "demographic bias." I discuss potential solutions that could be implemented in settings where this may be an issue.

8. Conclusion

This paper introduces a novel adaptive survey methodology that engages participants in stimuli creation and creates dynamic surveys that evolve with participant input. Applied first to issue salience, the CSAS method enabled participants to submit items reflecting prioritized issues. While traditional issue priorities like economic conditions, healthcare, crime, and education emerged as top-rated issues, the method also surfaced unique concerns such as candidate transparency and privacy protections. Given that participant-submitted questions frequently outranked those based on Gallup issue categories, CSAS could potentially identify public priorities that conventional survey approaches might ignore.

The method was also used in two additional settings: to study misinformation within the Latino community and measure salient concerns in local politics. In the Latino misinformation case, the analysis revealed that the most highly rated claims were partisan stereotypes, accurate statements, or widely reported allegations. In contrast, claims that were rated lower on accuracy were often objectively false and reflected more blatant misinformation. This application highlights CSAS's usefulness in identifying contested claims that might otherwise go unnoticed in traditional surveys. CSAS was also used to examine local political concerns, where its flexibility enabled the identification of both locally salient issues absent from national discourse and national issues not traditionally associated with local government like foreign policy.

A distinct advantage of the adaptive design is that the exploration of new items can yield a larger number of items than one would include in a traditional static survey. This not only makes it possible to explore new issues (e.g., artificial intelligence), but also to uncover unexpected areas of agreement (and disagreement) across partisan or ideological subgroups. Future studies could apply more sophisticated algorithms such as contextual bandits to account for subgroup heterogeneity and explore how question banks vary across social and political contexts (Offer-Westort *et al.* 2022). Although no substantial over-time sample imbalances were detected in this study, future applications could also use deconfounded Thompson sampling to adjust for potential variations in demographic composition over time (Qin and Russo 2022).[18]

Future research employing CSAS has the potential to explore a variety of topics where participant-generated content is particularly valuable. For instance, when voting in elections, voters not only consider issue positions or party affiliation, but also personal attributes such as honesty and competence and how politicians are perceived to navigate different economic and political conditions (Lenz 2012). CSAS could be useful in uncovering additional factors that influence voting decisions, but are not salient to researchers. Furthermore, when identifying norms, core beliefs, or key sources of identity within hard-to-reach communities, CSAS could reduce the misalignment of researcher-defined concepts with respondents' actual perceptions and interpretations of constructs (e.g., socially expected political behaviors or attitudes). Finally, when studying representation, scholars often rely on nationally salient issues, potentially overlooking the more idiosyncratic concerns of the public. CSAS could be used to complement existing measures of representation with dynamic question banks that include participant-generated issues.

[18]In settings where measures exhibit temporal variation, particularly in longitudinal settings, time-aware adaptive algorithms could account for temporal drift in estimates (Cavenaghi *et al.* 2021).

CSAS could also be used to develop measurement scales that more accurately reflect considerations that are important to participants. Given that many core constructs in the social sciences reflect latent variables (e.g., democracy, political sophistication, prejudice, identity), CSAS could help extract folk definitions and incorporate them into multi-item scales. Future research could also explore the feasibility of optimizing additional criteria, such as item discrimination and difficulty, to refine existing measures as new items are explored. By prioritizing the perspectives of study populations, CSAS could strengthen the connection between researchers and respondents and advance our understanding of public opinion and political behavior.

Acknowledgments. I would like to express my gratitude to the editor and three anonymous reviewers for their valuable comments. I am deeply indebted to Brandon Stewart, Don Green, Daniel Russo, Brendan Nyhan, Vin Arceneaux, Florent Parmentier, Jean-Philippe Cointet, and participants at the 2024 Exploring Misinformation Meeting at Sciences Po for their insightful feedback and suggestions. Special thanks are extended to Sciences Po CEVIPOF, the TIERED Project, and the Open Institute for Digital Transformations for their support.

Supplementary Material. For supplementary material accompanying this paper, please visit http://doi.org/10.1017/pan.2024.34.

References

Agrawal, S., and N. Goyal. 2017. "Near-Optimal Regret Bounds for Thompson Sampling." *Journal of the ACM* 64 (5): 1–24.

Anoll, A. P. 2018. "What Makes a Good Neighbor? Race, Place, and Norms of Political Participation." *American Political Science Review* 112 (3): 494–508.

Ansolabehere, S., and B. Schaffner. 2009. *Guide to the 2006 Cooperative Congressional Election Survey. Data Release No. 2*. Harvard University, *draft of February*, 9.

Cavenaghi, E., G. Sottocornola, F. Stella, and M. Zanker. 2021. "Non Stationary Multi-Armed Bandit: Empirical Evaluation of a New Concept Drift-Aware Algorithm." *Entropy* 23 (3): 380.

Chen, B., F. Yi, and D. Varró. 2023 "Prompting or Fine-Tuning? A Comparative Study of Large Language Models for Taxonomy Construction." In *2023 ACM/IEEE International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C)*, 588–596. IEEE.

Cortina, J., and B. Rottinghaus. 2022. "Conspiratorial Thinking in the Latino Community on the 2020 election." *Research & Politics* 9 (1): 20531680221083535.

Dillman, D. A., and L. M . Christian. 2005. "Survey Mode as a Source of Instability in Responses Across Surveys." *Field Methods* 17 (1): 30–52.

Dillman, D. A., J. D. Smyth, and L. M. Christian. 2014. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. Hoboken, NJ: John Wiley & Sons.

Egami, N., C. J. Fong, J. Grimmer, M. E. Roberts, and B. M. Stewart. 2018."How to Make Causal Inferences Using Texts." arXiv preprint arXiv:1802.02163.

Elliott, K. J. 2020. "Democracy's Pin Factory: Issue Specialization, the Division of Cognitive Labor, and Epistemic Performance." *American Journal of Political Science* 64 (2): 385–397.

Hopkins, D. J. 2018. *The Increasingly United States: How and Why American Political Behavior Nationalized*. Chicago, IL: University of Chicago Press.

Key, V. O. 1961. "Public Opinion and the Decay of Democracy." *The Virginia Quarterly Review* 37 (4): 481–494.

Krosnick, J. A. and, M. K. Berent. 1993. "Comparisons of Party Identification and Policy Preferences: The Impact of Survey Question Format." *American Journal of Political Science* 37 (3): 941–964.

Lenz, G. S. 2012. *Follow the Leader?: How Voters Respond to Politicians' Policies and Performance*. Chicago, IL: University of Chicago Press.

Lopez, J., and Hillygus, D. S. (2018). "Why So Serious?: Survey Trolls and Misinformation." Available at SSRN: https://ssrn.com/abstract=3131087 or https://doi.org/10.2139/ssrn.3131087.

Mellon, J., J. Bailey, R. Scott, J. Breckwoldt, M. Miori, and P. Schmedeman. 2024. "Do AIs Know What the Most Important Issue is? Using Language Models to Code Open-Text Social Survey Responses at Scale." *Research & Politics* 11 (1): 20531680241231468.

Montgomery, J. M., and J. Cutler. 2013. "Computerized Adaptive Testing for Public Opinion Surveys." *Political Analysis* 21 (2): 172–192.

Montgomery, J. M., and E. L. Rossiter. 2020. "So Many Questions, So Little Time: Integrating Adaptive Inventories into Public Opinion Research." *Journal of Survey Statistics and Methodology* 8 (4): 667–690.

Montgomery, J. M., and E. L. Rossiter. 2022. *Adaptive Inventories: A Practical Guide for Applied Researchers*. Cambridge: Cambridge University Press.

Offer-Westort, M., A. Coppock, and D. P. Green. 2021. "Adaptive Experimental Design: Prospects and Applications in Political Science." *American Journal of Political Science* 65 (4): 826–844.
Offer-Westort, M., L. R. Rosenzweig, and S. Athey. 2022. Battling the Coronavirus infodemic among Social Media Users in Africa. arXiv preprint arXiv:2212.13638.
Page, B. I., and R. Y. Shapiro. 1983. "Effects of Public Opinion on Policy." *American Political Science Review* 77 (1): 175–190.
Qin, C., and D. Russo. 2022. "Adaptive Experimentation in the Presence of Exogenous Nonstationary Variation." arXiv preprint arXiv:2202.09036.
Radford, A., K. Narasimhan, T. Salimans, and I. Sutskever. (2018). "Improving Language Understanding by Generative Pre-Training."
Rheault, L., and C. Cochrane. 2020 "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora." *Political Analysis* 28 (1): 112–133.
Ryan, T. J., and J. A. Ehlinger. 2023. "Issue Publics: How Electoral Constituencies Hide in Plain Sight." *Elements in Political Psychology*. Cambridge: Cambridge University Press.
Salganik, M. J., and K. E. Levy. 2015. "Wiki Surveys: Open and Quantifiable Social Data Collection." *PloS One* 10 (5): e0123483.
Stagnaro, M. N., J. Druckman, Berinsky, A. J., A. A. Arechar, R. Willer, and D. Rand. 2024. "Representativeness Versus Attentiveness: A Comparison Across Nine Online Survey Samples." Preprint, PsyArXiv.
Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I., 2017. "Attention is All You Need." In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS '17)*. 30 1–11.
Velez, Y. R., and P. Liu. 2024. "Confronting Core Issues: A critical assessmentof attitude polarization using tailored experiments." *American Political Science Review*. Advance online publication. https://doi.org/10.1017/S0003055424000819
Velez, Y. R. 2023. "Trade-Offs in Latino Politics: Exploring the Role of Deeply-Held Issue Positions Using a Dynamic Tailored Conjoint Method." *Aletheia*.
Velez, Y. R., E. Porter, and T. J. Wood. 2023. "Latino-Targeted Misinformation and the Power of Factual Corrections." *The Journal of Politics* 85 (2): 789–794.
Weller, S. C. 1998. "Structured Interviewing and Questionnaire Construction." In *Handbook of Methods in Cultural Anthropology*, edited by H. R. Bernard, 365–410. Walnut Creek, CA: AltaMira.
Wlezien, C. 2005. "On the Salience of Political Issues: The Problem with 'Most Important Problem.'" *Electoral Studies* 24 (4): 555–579.

Cite this article: Velez, Y. R. (2025). Crowdsourced Adaptive Surveys. *Political Analysis*, 1–14. https://doi.org/10.1017/pan.2024.34