Centre for Social Informatics

Faculty of Social Sciences

University of Ljubljana

The Price of Precision

A Qualitative Study for Cost-Error Balancing in the Decision-Making Process when Deciding on Probability-Based vs Nonprobability-Based Online Survey Panels

Andrea Ivanovska, Vasja Vehovar, Michael Bosnjak, Bojana Lobe

Ljubljana, March 2025

Table of Contents

1 Introduction

Over the last three decades, the landscape of survey research has been fundamentally changed by online survey mode. In the context of general population surveys, the online panels emerged and become the prevailing survey data collection format. With some exceptions, such as complex interaction surveys (e.g. the Programme for the International Assessment of Adult Competencies; Maehler et al., 2025) or essential official statistics survey (e.g. Labour Force Survey; Eurostat, n.d.) addressing factual and behaviour topics, they almost entirely replaced traditional surveys (face-to-face, telephone, mail) when attitudes are targeted. Online panels consist of a group of registered respondents who have preliminarily agreed – against some incentives – to regularly participate in online surveys. These panels effectively streamline and optimise the data collection process (Callegaro et al., 2015) and can be classified into probability-based online panels (PBOPs) and non-probability-based online panels (NPBOPs) (Cornesse & Blom, 2020). PBOPs recruit members with probability-based sampling design, relying on offline methods for the initial recruitment (such as personal visit or postal mail) to ensure that every member of the target population has a known, non-zero probability of selection (Bosch & Maslovskaya, 2023; Callegaro et al., 2015). In contrast, NPBOPs, also known as opt-in panels, rely on a self-selection process where individuals volunteer to join, often drawn by advertisements promising monetary rewards for their participation (Cornesse & Blom, 2023; Slamowicz et al., 2026). Because these respondents are primarily motivated by financial incentives, they frequently exhibit higher rates of satisficing behaviors, such as straight-lining, which introduces considerable measurement error into the survey data (Cornesse & Blom, 2023; Slamowicz et al., 2026).

This introduces an important dilemma for researchers and survey practitioners, who need to decide between the statistical rigor of PBOPs and the resource efficiency of NPBOPs. Today, business, marketing, and public opinion (i.e., polls) research rely almost exclusively on NPBOPs, while national statistical offices continue to adhere strictly to probability sampling principles. The remaining academic, government, and nonprofit sectors are divided between these two options, with unclear rules and guidance on how to decide. The related dilemmas are central focus of our research.

To address this gap and clarify the decision-making landscape, this study pursues two core objectives First, we will identify the specific attributes and characteristics which decision-makers use to assess probability and non-probability panels. Second, we aim to understand how these different attributes are weighed against each other in practice. By exploring whether survey practitioners rely on compensatory versus non-compensatory decision methodologies, or use specific cognitive heuristics, we intend to map the real-world logic that drives these complex methodological choices.

The specific research questions are as follows:

RQ1: How do applied researchers (i.e. in the role of research manager) cognitively frame and operationally navigate the trade-offs between the statistical rigor of PBOPs and the resource efficiency of NPBOPs?

RQ2: What heuristic mechanisms, institutional pressures, and professional norms define the perceived "acceptability" of survey error in practice?

This study uses an exploratory, qualitative approach with interactive vignettes to answer these questions. By interviewing survey researchers and practitioners from academic, government, and nonprofit sectors, we aim to identify the specific attributes that drive these methodological trade-offs, such as costs and bias, for example.

1.1 Background

Theoretical Foundations: Probability vs. Non-Probability Panels

PBOPs recruit members with probability-based sampling design, relying on offline methods for the initial recruitment (such as personal visit or postal mail) to ensure that every member of the target population has a known, non-zero probability of selection (Bosch & Maslovskaya, 2023; Callegaro et al., 2015). In contrast, NPBOPs (also known as access, opt-in, or volunteer panels) rely on a self-selection process where individuals volunteer to join the panel (Vehovar et al., 2016). This results in an unknown probability of selection and therefore an unknown bias of the estimates can appear. On the other hand, the primary strength of these panels is speed of data collection and considerably lower costs compared to than PBOS or face-to-face alternatives (Callegaro et al., 2015; Lehdonvirta et al., 2021).

Empirical evaluations that compare PBOPs and NPBOPs to census data or data from high quality governmental surveys (referred to as benchmarks) demonstrate a clear hierarchy of accuracy. In a comprehensive review of the literature, Cornesse et al. (2020) concluded that probability-based surveys in general (some PBOPs were a subset of the analysed studies) consistently resulted in significantly higher accuracy than non-probability sample surveys, demonstrating that probability sampling designs outperform non-probability designs (which are predominantly online access panels) across various countries and topics. Furthermore, MacInnis et al. (2018) found that probability samples were consistently the most accurate, a trend that remained stable over time despite declining response rates. Mercer and Lau (2023) quantified the difference in quality, reporting that opt-in samples had an average absolute error of 5.8 percentage points across benchmarks, which was more than double the 2.6 percentage point error of probability-based panels. The bias in NPBOPs is often widespread and substantial, particularly concentrated within specific subgroups and rare characteristics.

MISSING: Empirical evidence on costs, a summary of the empirical evidence specifically regarding the costs financial differences/cost ratios between PBOPs and NPBOPs.

1.2 Additional attributes influencing decision-makers

Traditionally, the evaluation of survey errors has been grounded in the Total Survey Error (TSE) framework (Biemer & Lyberg, 2003; Lyberg & Stukel, 2017), which prioritises the minimisation of bias and variance of the estimates to achieve statistical accuracy. Very often, decisions regarding survey design are guided exclusively by error-related aspects, e.g. highest response rate or lowest bias of the estimates (Vehovar and Beullens, 2018).

However, survey practitioners operate within the much broader constraints of Total Survey Quality (TSQ; Lyberg & Stukel, 2017; Eurostat, 2021). In this framework, accuracy is just one dimension alongside other indicators (Callegaro et al., 2015). According to Eurostat (2021), quality also encompasses relevance (the degree to which statistics meet user needs), timeliness and punctuality (the time lag between the event and the release of data), accessibility and clarity (the ease of access and availability of metadata), and coherence and comparability (the adequacy of statistics to be reliably combined or compared). These additional indicators are non-statistical in nature. By addressing these specific managerial and contextual aspects, the framework ensures a holistic assessment of the survey's fitness for use.

To fully grasp the dynamics of organizational decision-making, we also have to consider the impact of non-statistical factors such as institutional pressures, regulations, habits, and professional reputation. Sociological institutionalism suggests that actors within organizations are heavily driven by a desire for legitimacy, often prioritizing social and cultural conformity over purely rational or efficient performance (Miller, 2005). Formal regulations and laws exert coercive pressures that mandate specific behaviors from organizations (Economou et al., 2023). However, when rules are unclear or the environment is highly uncertain, decision-makers often navigate this by simply imitating the practices of peers they view as successful. Over time, these emulated actions can quickly solidify into ingrained habits and routines, creating a path dependence that leads to organizational inertia and restricts the exploration of alternative choices.

Furthermore, the need to protect professional reputation acts as a powerful catalyst for this institutional conformity (Perkins & Shortland, 2022). Because decision-makers face heavy public scrutiny and worry about damaging their reputations, they often focus on avoiding risk. As a result, they tend to rely on standard, boilerplate solutions rather than creating custom strategies that fit their specific needs.

1.3 The decision-making process

Ideally, survey design is conceptualised as a rational optimisation problem. In this view, researchers would weigh the marginal cost of bias reduction against the marginal utility of data accuracy, aiming to minimise the MSE within a fixed budget (Groves, 1989). Standard economic theory assumes individuals are perfectly rational, but real-world decision-makers face severe limits regarding memory, attention, and information-processing capacity (Schwarz et al., 2022). However, decision science suggests that this level of optimisation is rarely achievable in practice. According to the theory of bounded rationality (Simon, 1955;

1990), individuals are limited by the information they have, the cognitive limitations of their minds, and the finite amount of time they have to make a decision. Therefore, the decision process shifts from strict calculation to the use of cognitive shortcuts or “heuristics” (Tversky & Kahneman, 1974). Instead of optimising (finding the absolute best solution), decision-makers often search through alternatives until they find one that meets an acceptable threshold of quality (Simon, 1990; Schwarz et al., 2022).

In the context of survey methodology, researchers navigate complex choices between panel types by using different decision-making strategies that range from exhaustive optimisation to cognitive shortcuts. These strategies can be broadly categorized into compensatory and non-compensatory models, which differ significantly in cognitive demand and the thoroughness of the evaluation.

1.3.1 Compensatory decision-making models

Compensatory models involve a comprehensive evaluation where a high value on one attribute can offset, or "compensate" for, a low value on another. These are typically more thorough and demanding cognitive strategies.

Additive linear strategy: This is a rigorous approach where every attribute is weighted by its importance and summed to calculate a global utility for each alternative. While this method is most likely to identify the uniquely "best all-around" option, it requires substantial time and cognitive effort.

Additive difference strategy: Decision-makers compare two alternatives at a time, calculating differences attribute-by-attribute and summing these differences to determine a "winner" for that pair. The process continues through pairwise comparisons until a final choice is reached.

1.3.2 Non-compensatory decision-making models

Non-compensatory strategies are often "unforgiving" and do not allow trade-offs; if an alternative fails to meet a specific threshold on a single key attribute, it is immediately rejected regardless of its other strengths. These models are generally less cognitively demanding and yield faster answers.

Satisficing (conjunctive): The decision-maker sets minimum "acceptability cutoff" points for all important attributes. They then select the first alternative that meets these thresholds across the board.

Lexicographic strategy: The researcher identifies the single most important attribute and chooses the alternative that performs best on that specific dimension. If there are ties, they proceed to the next most important attribute.

Elimination by Aspects: Decision-makers select a salient attribute and set a cutoff point; any alternative falling below that cutoff is eliminated. This process is repeated with subsequent attributes until only one option remains.

Dominance: This involves searching for an alternative that is at least as good as all others on every attribute. It is often used as an initial stage to "weed out" the weakest options from a large choice set.

While compensatory models are more likely to identify optimal solutions, non-compensatory, attribute-based strategies are more "cognitively frugal" and adaptive when information is unreliable or decisions must be made quickly.

Rather than basing their decisions on complicated cost-error calculations for every project, researchers may depend on other cues, such as personal beliefs about acceptable quality, or social cues about what is considered as credible in social research. Understanding these cues and mechanisms is important, as they explain why decisions often deviate from the technical predictions of the TSE framework.

2 Methodology

2.1 The aim of the study

This study relies on a qualitative exploratory design to provide insight into the decision process researchers undertake when choosing between a sample from a PBOP and a NPBOP. The objective is to explore how researchers cognitively and operationally navigate trade-offs between data quality and resource limits. Specifically, we aim to identify whether practitioners employ compensatory strategies, where high accuracy can offset high costs, or non-compensatory strategies, where specific attributes like a budget cap or a methodological mandate act as absolute "deal-breakers".

The qualitative research determined which attributes are important for decision-making regarding the above dilemma, reflecting realistic variability in professional practice. The preliminary research identified the following list of attributes, which is, of course, not closed or final: the absolute cost of the survey, the share of survey cost in relation to the total project budget, the magnitude of additional bias in NPBOP, the share of variables that are biased, and the project purpose (exploratory or policy creation). The specific levels for each attribute were also of our interest, such as cost ratios between the PBOP and NPBOP, or the levels of the relative bias defined as the proportional difference from the true value, rather than absolute percentage points (e.g. below 5%, 5-10%, 10-30% etc).

While the tension between statistical rigor (TSE) and practical constraints (TSQ) is well-documented (Lyberg & Stukel, 2017), and decision theory explains why individuals satisfice rather than optimise (Simon, 1990), we lack empirical evidence on how this works specifically in survey design. Do researchers prioritise absolute cost or cost ratios? Is acceptable error defined by statistical power or by industry standards? To understand when researchers decide to abandon the gold standard for a cheaper alternative, we cannot rely on theoretical models alone; we must investigate the lived experience of the decision-makers themselves. This study bridges that gap by exploring the specific attributes and trade-offs that drive these methodological choices.

2.2 Participants and Sampling

Data were collected through 22 in-depth semi-structured interviews with research and survey research practitioners. The sample was selected using purposive sampling to ensure participants possessed relevant professional experience in designing, budgeting, or managing survey projects. The participants predominantly included researchers and project managers from the academic sector, but also from government, and nonprofit sectors. The business sector is not included, because they almost exclusively rely on NPBOP.

The interviews ranged from 13 to 51 minutes, with an average of 30 minutes. 19 interviews were carried out online, while 3 were done in person. The reliance on video-based formats for the majority of interviews aligns with methodological assessments that view video interviewing as the closest alternative to in-person interaction, aiming to capture the spontaneity of traditional face-to-face exchanges (Lobe et al., 2022).

Sample description table MISSING:

- Sector (academic, government, nonprofit)
- Role (researcher responsible for research design, researcher in charge of survey process)
- Experience in relation to here discussed decision (hypothetical answering, little experience, a lot of experience)
- Country (Slovenia, other)

2.3 Data Collection

Participants were presented with hypothetical research scenarios requiring a choice between a PBOP (characterised by higher cost and lower bias) and a NPBOP (characterised by lower cost and potential bias). Unlike standard vignette studies where the stimulus remains constant, this study employed an iterative design based on theoretical sampling. The vignettes were not static; they were dynamic instruments refined throughout the data collection process to test emerging hypotheses and maximise the discriminatory power of the attributes. The changes made to the vignette, within one interview session and between successive interviews, were based on three primary streams:

- Level Calibration (Testing Compensatory Limits): The values within the vignettes (e.g., price differences between €4,000 and €8,000 vs. €4,000 and €40,000) were manipulated to probe the participants' "breaking points." This allowed us to observe the limits of compensatory logic, the specific threshold where a panel's statistical rigor can no longer justify its cost.

- Attribute Refinement (Testing Framing Effects): Vignettes were modified to isolate cognitive biases, such as the gain-loss framing effect. For example, we refined cost attributes to distinguish between a "savings" frame (where the surplus from a cheaper option is available for discretionary use) and a "loss" frame (where higher costs force a direct reduction in other project tasks), observing how these frames shift the participant's reference point.

- Attribute Inclusion (Identifying Non-Compensatory Cutoffs): New attributes were introduced to identify non-compensatory constraints. For example, after early participants noted that certain institutional norms or client requirements were non-negotiable, we included "Institutional Mandate" as an attribute. This allowed us to see if such factors functioned as lexicographic triggers, where the researcher chooses based on a single most important attribute regardless of other benefits.

Data collection took place between February and November 2024. Interviews were conducted in either a one-on-one or two-on-one format (two interviewers, one participant). Given that the participants were professional peers, the data collection process required managing a 'fluid positionality' (Buys et al., 2022). This involved balancing the 'insider' status of shared technical language and methodological expertise with the 'outsider' role required to objectively probe the participants' specific decision thresholds without assuming shared meaning (Buys et al., 2022).

2.4 Data Processing

Of the 22 interviews, three were conducted in English, while 19 were conducted in Slovenian. The data processing followed a two-stage protocol to ensure analytical consistency across languages:

- Transcription: All audio recordings were transcribed using a denaturalised (intelligent verbatim) transcription protocol. This method retained the full semantic content of the responses while removing non-verbal utterances (e.g., stutters, pauses, laughter) and prosodic features. This approach was selected because the analysis focused on the logic and content of the decision-making process rather than the linguistic characteristics.

- Translation: The 19 Slovenian transcripts were translated into English to facilitate a unified analysis. A meaning-based translation strategy was used, prioritising the conceptual equivalence of technical survey methodology terminology (e.g., “representativeness,” “non-response bias,” “weighting”) over literal word-for-word translation.

2.5 Data Analysis

The data were analysed using Reflexive Thematic Analysis (TA; Braun & Clarke, 2006; 2019). This approach was selected to acknowledge the active role of the researcher in knowledge production, moving beyond a simple description of the decision-making process to an interpretation of the latent meaning structures. The analysis did not seek to quantify the frequency of choices but rather to discover the nuances in researchers' decision-making.

The analysis adopts a critical realist stance. It assumed that while budget constraints and error rates are real (ontological realism), the interpretation of these constraints, such as the acceptability of error, is socially constructed (epistemological constructivism).

Coding was done using an abductive coding strategy, involving a continuous iterative dialogue between the data and existing theoretical frameworks (Tavory & Timmermans, 2014). This approach allowed for the integration of established survey methodology concepts with new insights. The TSE and TSQ frameworks served as “sensitising concepts” (Blumer, 1954). Initial codes were informed by these frameworks to identify explicit references to technical constraints, such as coverage error, or relevance. This ensured the analysis remained grounded in the material reality of statistical rigor. The codes were also expanded to include cognitive categories from decision science. We specifically coded for indicators of satisficing (where respondents sought a "good enough" threshold) and additive linear logic (where respondents performed exhaustive trade-offs between bias and cost). This process allowed for the identification of nuances standard frameworks could not explain, such as the use of non-compensatory heuristics as a defensive shield against professional criticism.

Codes were clustered into candidate themes through a process of reflexive engagement. The final thematic structure was not only determined by the frequency of codes, but also their ability to explain the latent tensions in the decision-making process. We prioritised themes that captured the relationships between the strict constraints of survey error (derived from the TSE framework) and the “soft” social strategies researchers use to navigate them (derived inductively). This process resulted in three overarching themes: the financial psychology of decision-making, the social construction of accuracy, and the management of professional risk. These themes map directly onto our research inquiries. Theme 1 and Theme 2 primarily address RQ1, illustrating how researchers navigate trade-offs through grant-based financial logic and cognitive heuristics like satisficing. Theme 3 and components of Theme 2 address RQ2, highlighting how the “acceptability” of error is socially constructed through professional norms, reputational defense mechanisms, and institutional constraints.

3 Results

3.1 Theme 1: The financial psychology of decision-making

The first dimension of researchers' decision making process is, inevitably, economic. However, the analysis of the qualitative data suggests that it does not conform to the cost-error optimisation models often assumed in survey design literature, where researchers are expected to rationally minimise error for a fixed cost (Groves, 1989). Instead of this mathematical optimisation, decision-making is driven by a grant-based rationality, where the relative proportion of the budget and the bureaucratic expectations outweigh methodological precision.

3.1.1 Subtheme 1.1: Relative cost

When researchers evaluate the trade-off between a high-cost PBOP (e.g., €8000 or €40000) and a low-cost NPBOP (e.g., €4000), they rarely assess the absolute value of the money. Instead, they consistently applied a 'cost-ratio' logic, where the acceptability of the cost was determined by its percentage of the total project budget.

If the cost of the survey was a small fraction of the total grant (often cited as under 10%), the decision-making process was bypassed entirely, and the "gold standard" was selected automatically. Participant 2015 explained, "If the price is [double], I would go for probability… when we look at it within a relatively large project, which is three hundred grand, it seems quite easy to justify those four grand for empirical research". Similarly, Participant 2018 agreed, noting that "spending ten percent on, on data collection seems minor to me. So I would go for the probability sample in any case".

This suggests a lexicographic strategy , where the "cost ratio" attribute is considered first; if it falls below the 10% threshold, the decision is made immediately without considering other attributes.

However, a point where the absolute cost becomes important exists. The data suggests that when the total project budget shrinks (e.g., to €100000), or when the cost of the survey exceeds a certain psychological threshold (e.g., €16000 or €40000), the calculation shifts dramatically. Participant 1001 elaborated, “So again, if you push it down to 100000, it'll be entirely different”. Participant 2014 explained, “I don't know, the price of 16,000 seems really excessive to me and I would probably try to negotiate. In the end…I would probably choose the non-probability one.”.

3.1.2 Subtheme 1.2: Saving opportunities

The analysis identified that researchers don’t weigh the opportunity cost of spending money on data vs. other research activities equally for all types of savings. When the cost of precision threatened financial compensation, data quality was instantly sacrificed. Participant 2007 reinforced this boundary, stating they would push for the more expensive PBOP only if it didn't endanger the team: “In that case, I would try in every way to squeeze out the 40000, if that didn't affect salaries, or urgent training”. Here, "staff salary" acts as a non-compensatory cutoff point ; no amount of data precision can compensate for the failure to meet this basic financial requirement.

On the other hand, researchers rejected trading data quality for personal benefits (e.g., conferences, professional trainings). Participant 2008 elaborated that professional development funds are not factors “for which I would want to save money,” but if the savings allowed for “an additional three-month employment of a person... that would write an article,” the choice changes.

3.1.3 Subtheme 1.3 The “good manager” paradox

A counter-intuitive finding was that the drive to be a “good manager” led to contradictory decision-making depending on how the researcher defined administrative competence. The analysis revealed a paradox where researchers sometimes selected the expensive PBOP specifically to maximize budget execution, while others rejected it to demonstrate pragmatic realism. In both cases, the methodological choice was secondary to the desire to satisfy a specific administrative identity. In many institutional contexts, especially in public universities and government agencies, returning unspent grant money is viewed as a failure of planning or a signal of limited capacity, potentially jeopardising future funding.

This creates an incentive to maximise expenditure rather than efficiency. Participant 2011 explicitly stated, "basically I got the money from the state to do this... I know how to spend 40 grand for such a thing too". Furthermore, Participant 2005 stated they would not split the budget: "This is my approach to research: better to do one thing very properly... than several um, tiny, 'worse' projects".

However, for other researchers, the definition of a "good manager" was inverted. Instead of budget execution, they prioritised resource realism. Some researchers contradicted this, explaining that "if you wanted to conduct every research really perfectly, then anyway we always put too little money into it, right? You always have to make a compromise because you always have too little money or, or people, or anything" (Participant 2003). They took on a role of a manager that is good because they are realistic, and accept that they will always have to make trade-offs: "The main factor would be what I want to achieve and how much of that I can compensate [compromise on], right. If it is very important, I will just try to go for the best possible. Otherwise, I will maybe compensate [compromise] a little".

3.2 Theme 2: The social construction of accuracy

This theme explores how researchers cognitively align their financial and professional realities with their scientific ideals. It reveals a tension between sticking to a "gold standard" ideology and using "fit for purpose" justifications and personal thresholds.

3.2.1 Subtheme 2.1: The "gold standard" ideology

The analysis revealed that probability sampling is the gold standard among most researchers, and that it exists beyond economic critique. Participant 2006 explained: "Here it isn't cost-benefit, but just benefit... we cannot deviate from the standard". Participant 2016 even exclaimed that without the possibility of a probability online panel, "we would have exited the project".

This stance is rooted in a fundamental skepticism of non-probability methods. Participant 2018 compared successful NPBOP results to luck: "Meaning, probability sampling is simply the only one that systematically and in principle gives quality data... Otherwise, of course, here and there we can win the lottery". The same participant described a situation where they would discard free or new data for a secondary analysis because it fails the fundamental sampling assumption. Participant 2019 shared a similar view, stating “But if I had some money, I would throw everything into the sample down to the last cent.”

3.2.2 Subtheme 2.2: The ethical importance of coverage

Beyond statistical rigour, a major driver for the “gold standard” is the ethical dimension of coverage error. Researchers contextualise the exclusion of offline populations (common in NPBOPs) not just as a statistical flaw but as a moral failure. Participant 2005 said “The aspect of non-users is being neglected... it would be very important to say, ok, now we need a very unbiased estimate,” and Participant 2009 explained: “If you are missing a whole vulnerable group or if variables are, um, biased. Yes, on a national estimate it is the same, right? But in reality, in most cases, we are interested in differences between, say, one vulnerable group and [another] one.” In the Total Survey Quality (TSQ) framework, Completeness and Relevance are key dimensions. If a survey aims to inform public policy on aging or poverty, a sample that systematically excludes the digital disconnected is unfit for use. Researchers pay for PBOPs (which often include offline recruitment or mixed modes) to ensure that the invisible populations are included. Researchers viewed the production of biased estimates for vulnerable populations as morally “unethical.” Participant 2018 even argued that “replacing a probability sample with a non-probability one is a question of ethics” if it leads to poor quality research that wastes participants' time.

3.2.3 Subtheme 2.3: “Fit for purpose” and negotiating quality

Despite the gold standard ideology, negotiating quality was common. Participants acknowledged that “Research design is often a negotiation with reality” (Participant 2002). Contextualisation of bias was heavily dependent on the “fit for purpose” principle. Participants distinguished clearly between high-stakes policy research and exploratory studies. As Participant 1001 noted, bias “is not just a number; it is a public health risk” depending on the topic. However, if the research involved “unimportant things” or was merely for correlation or modelling, researchers were more willing to compromise. As Participant 2002 explained, “In modelling, getting the direction and strength right is enough, precise numbers don't matter.”

This pragmatism led to compensatory strategies, such as increasing sample size in non-probability panels to “drown out” error. Participant 2010 argued that sample size outweighs fundamental bias if it means slightly larger subgroups, though this was rejected by others (Participant 2018) who get “red spots on my face” when researchers claim representativeness based solely on sample size. This represents a compensatory model , specifically an additive linear logic , where researchers weigh the marginal utility of a larger sample size against the known bias of the non-probability mode.

3.2.4 Subtheme 2.4 Manageable bias

Interestingly, the evaluation of bias was often led by personal thresholds rather than statistical power calculations. Participants cited arbitrary “mental cut-offs” for acceptable error, most commonly around 5-10% relative bias. Participant 1001 stated “ So my personal threshold is 5%”, while Participant 2006 noted that even low numbers can provoke skepticism: “You know what, five percent is the point where you raise your eyebrows a little. The share is surely still statistically significant. Five percent means four versus four point two”. Participant 2001 described this comfort zone explicitly: “Well, if it were all under ten percent, well, then I would sleep relatively well, right. ... With 10 percent I would then, would then say, okay, this is now a note, it works... So five or ten, I think, in the goal.” These "mental cut-offs" function as conjunctive satisficing rules , where alternatives are rejected if they do not meet a predetermined aspiration level for accuracy.

Conversely, very high levels of bias (30% or more) were universally rejected, often viewed as structurally compromising the research regardless of the “fit for purpose” argument. Participant 2010 explained: “If it were only point estimates, probably thirty percent is too much. If I were interested in correlation, I'd take even thirty percent. But there, more than thirty, then no, no. That is just too big an error.”

Participants also had personal thresholds about the amount of variables that could have a certain amount of bias, regardless of its magnitude: “I mean, thirty percent is a lot of bias... but a tenth [of all variables] is actually very little. I don't know, if it's twenty questions, that's two, right. So maybe I would still stick with the non-probability one.” Participant 2012 explained: “Then I would definitely think about it a bit more [if half of the variables are biased], about moving away from this non-probability online panel toward the probability sample. Yes, that would probably be the most important factor in the decision". Participant 2013 identified a specific threshold between 10-20% where "the amount of bad data becomes unacceptable."

3.3 Theme 3: Managing professional risk and uncertainty

While economic factors and methodological considerations define the boundaries of possibility, social factors dictate the choices made within those boundaries. The analysis suggests that bias is primarily contextualised as a professional risk, managed through strategies of blame avoidance and reliance on social trust.

3.3.1 Subtheme 3.1: Shielding the researcher

The data is filled with examples of researchers using probability sampling methodology as a defensive shield. Participant 2005 confessed to an "inferiority complex," fearing rejection from editors based on their location, leading to a belief that they "simply have to perhaps do even more proper research to get somewhere". Participant 2006 stated, "We are already constantly under fire... we at least have to have some backup". Similarly, Participant 2001 admitted, "it doesn't seem sensible to me to open myself up then, for someone to reproach me... why I didn't choose a better sample". Choosing the "gold standard" to avoid reproach is a non-compensatory dominance strategy used to weed out any alternative that could be perceived as a "loser" during the peer-review process.

It is important for participants to be able to back up their methodological decisions before stakeholders (Participant 2015) and in the publication process (Participant 2004), where a "decent study" is "basically much more publishable". Although some participants noted that journals are "quite flexible" regarding NPBOPs (Participant 2003), they emphasise that "reporting should be very careful" (Participant 1002) to mitigate this risk.

3.3.2 Subtheme 3.2: External pressure and institutional constraints

External pressure played a decisive role. For national surveys or projects funded by specific agencies, the methodology was often dictated from above, leaving the researcher no room for trade-offs. "These institutions decide…based on an expected bias, which road to travel down", said Participant 1001. Participant 2014 highlighted specific challenges in the public sector, where the requirement for competitive bids forces the selection of cheaper options: "But because public funds are involved, it needs to be justified and basically have three offers, right". The situation gets more complicated in international projects, as Participant 2014 explained, "If I were doing, say, international research with cooperation with some foreign partners, my decision would most likely also be influenced by what the possibilities for sample selection are in other countries".

3.3.3 Subtheme 3.3: Reducing complexity through trust

Researchers frequently described the methodology of commercial panels as lacking transparency. Participant 1002 notes the evasion of vendors when asked about technical details: "You ask them basic question, can you produce the weights? They say, no, we are fine. All your sample is very representative of the population. You don't need to use weights. And this is a little bit concerning".

To solve these issues, researchers rely on trust and social cues. Following Granovetter's (1985) theory that economic actions are embedded in structures of social relations, researchers substitute the missing technical information with relational trust. The decision then relies on the vendor's position in the academic network as a proxy for quality. Participant 2014 chose NPBOP vendors because they "collaborate with academic institutions, it gives them credibility". This reliance on trust operationalises the trade-off by reducing complexity. Instead of conducting a rigorous TSE assessment of every vendor (which is impossible due to lack of transparency), researchers use relational clues: if a vendor is trusted by the community, their data is accepted as precise enough. Although, this scepticism can sometimes be more deeply rooted, as seen in one researcher who rejected the idea of panels for data collection altogether based on the current market situation: "Well, people probably know [non-probability panel agencies], and similar. Now, if someone comes new making a probability one, I mean, how to assure people…that we are not the same." This sentiment highlights how trust serves as a cognitive shortcut to reduce the complexity of the methodological choice. This demonstrates a recognition heuristic , where researchers rely on name recognition and academic reputation to make a choice under conditions of low or unreliable information.

These three themes reveal that the decision-making process is governed by a specific set of heuristic rules. Table 1 summarises the key attributes identified within each theme and the prevailing consensus among researchers about their acceptable thresholds.

Table 1: Summary of decision-making attributes and observed researcher consensus

Theme	Key decision attribute	Decision Logic (Theoretical Match)
1. Financial psychology	Cost ratio (vs. absolute price)	Lexicographic Strategy: Reviewing the attributes and picking the one most important attribute (e.g., <10% cost ratio), then choosing the best alternative on that attribute.
	Budget flexibility	Satisficing Cutoff: Setting an "acceptability cutoff" point (e.g., salary protection) and looking for the first alternative that is at least as good as that value.
2. Social construction of accuracy	Relative bias threshold	Non-Compensatory Rejection: An "unforgiving" strategy where if an alternative is over a threshold (e.g., >30% bias), it is rejected regardless of other features.
	Scope of bias	Satisficing (Conjunctive): A strategy used to screen out "losers" by setting cutoff points on all important attributes (e.g., 10–20% of biased variables).
	Research purpose ("fit for purpose")	Compensatory (Additive Linear): A rigorous strategy that weights all attributes by importance to find the uniquely "best all-around" alternative.
	Coverage error	Protected Value (Non-Compensatory): Treating an aspect as an absolute requirement (e.g., ethical inclusion) that cannot be traded off for cognitive or financial savings.
3. Professional risk	Reputation	Dominance / Blame Avoidance: Searching for an alternative (e.g., PBOP) that is at least as good as every other alternative on all important attributes to avoid "rash steps".
	Vendor trust	Recognition Heuristic: Choosing the first alternative that is recognized (e.g., trusted academic collaborator) when information about other alternatives is low.

4 Discussion

The purpose of this study was to explore the “price of precision” and the complex decision-making process researchers go through when choosing between the statistical rigor of PBOPs and the resource efficiency of NPBOPs. While the TSE framework prescribes a mathematical minimisation of bias (Biemer & Lyberg, 2003), our findings suggest that in practice, this decision is rarely a calculation. Instead, researchers describe it as a negotiation shaped by limited rationality, budget constraints, and professional concerns.

The most significant finding of this study is the empirical confirmation of “satisficing” in survey design. In this process, the decision-maker does not thoroughly consider every option in order to determine the most optimal option. Instead, they operate using an “aspiration level” (Simon, 1955) as a minimum threshold of acceptability based on their goals and constraints. Researchers then search through methodological alternatives sequentially and choose the first one that meets or exceeds this aspiration level. As predicted by Simon (1955; 1990), researchers settle for designs that meet acceptable quality standards rather than striving for a theoretical maximum of accuracy.

The “5-10% relative bias threshold” that our analysis identified is the perfect example of this mechanism. This figure does not appear to be derived from specific power calculations for the projects in question, but could be related to the ingrained and often misunderstood statistical convention of $p<.05$ (Eden & Inan, 2022). This reliance on a fixed percentage acts as a cognitive 'anchor' (Tversky & Kahneman, 1974), allowing researchers to bypass the complex cognitive load of estimating bias for every specific variable. Survey design is thus often driven by following established norms (the “gold standard”) rather than a substantive calculation of outcomes (Schwarz et al., 2022). This mirrors the most recent methodological guidelines. Rohr et al. (2024) explicitly conclude that non-probability surveys should be restricted to specific non-inferential purposes, including exploratory studies for hypothesis generation, survey experiments prioritizing internal validity, and studies of hard-to-reach populations.

Contrary to the economic assumption that researchers try to maximise data quality for the lowest cost, our results revealed decision making relative to the grant. The decision to use a PBOP is often determined not by its absolute price, but by its ratio to the total budget. This cost-ratio creates a paradox where the same absolute cost for data quality can be too expensive in a small project but negligible in a large one, despite the scientific value of the data remaining constant.

Furthermore, the "good manager paradox" highlights a misalignment between administrative incentives and methodological efficiency. In the academic and government sectors, the pressure to spend grants can facilitate the selection of more expensive probability samples. This suggests that the "price of precision" is sometimes paid to satisfy bureaucratic requirements of budget execution, and not to reduce error. This aligns with recent findings in design science by Nickel et al. (2024), who concluded that design trade-offs are rarely purely technical but are shaped by conflicting stakeholder goals and organizational constraints. Similarly, we found that some researchers feel administrative pressures to execute budgets often override the theoretical freedom to optimise for survey error. Facing these conflicting stakeholder pressures, researchers ultimately treat methodology less as a scientific instrument and more as a tool for risk management. In this context, the "gold standard" of probability sampling serves primarily as an insurance policy for the researchers' reputation, functioning as a mechanism of blame avoidance (Hood, 2011) where adherence to established norms protects the practitioner from scrutiny.

4.1 Strengths and limitations

According to the evaluative criteria for qualitative research established by Lincoln and Guba (1986), this study prioritised trustworthiness over the positivist standards of generalisability and reliability. Credibility was ensured through the researchers' "fluid positionality", where shared professional expertise with participants facilitated deep technical discussions and minimised misunderstandings. Dependability and confirmability were maintained through an audit trail of the iterative vignette design and additional notes (Ahmed, 2024), which helped manage researcher bias and documented the evolution of the decision-making attributes. Regarding transferability, the sample was heavily skewed toward academic and government researchers. Private sector researchers would operate under different profit-maximisation constraints. The specific financial thresholds (e.g., €40000) may not transfer to the commercial sector, but we have no reason to assume different underlying decision-making mechanisms.

An additional limitation of this study is its reliance on hypothetical vignettes (Erfanian et al., 2020). While vignettes allow for the isolation of specific decision attributes (such as cost ratios), they cannot fully replicate the high-stakes pressure of spending actual grant money.

4.2 Future research

Building on these exploratory findings, future research should move from identifying attributes to quantifying their influence through experimental designs. Following the methodological framework for attribute development outlined by Coast et al. (2012), the qualitative themes isolated in this study (e.g., cost ratios, bias magnitude, funding source) can be operationalised into precise experimental levels. A vignette-based conjoint analysis or discrete choice experiment could then be deployed to calculate the relative importance of each attribute and model the exact cut-off points where professional preference shifts from statistical rigor to economic efficiency.

5 References

Maehler, D. B., Hernández-Torrano, D., Courtney, M. G., Fischer, F. P., Hollricher, L. F., & Gorges, J. (2025). PIAAC Survey of Adult Skills: A review of the research landscape. *International Review of Education*, *71*(4), 685-714.

Eurostat. (n.d.). *EU Labour Force Survey - new methodology from 2021 onwards*. Statistics Explained. Retrieved March 2, 2026, from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=EU_Labour_Force_Survey_-_new_methodology_from_2021_onwards

Cornesse, C., & Blom, A. G. (2023). Response quality in nonprobability and probability-based online panels. Sociological Methods & Research, 52(2), 879–908. https://doi.org/10.1177/0049124120914940

Slamowicz, S., Pennay, D., Neiger, D., & Phillips, B. (2026). Reducing the bias from probability and nonprobability online panels by excluding satisficers. Journal of Survey Statistics and Methodology.

Ahmed, S. K. (2024). The pillars of trustworthiness in qualitative research. *Journal of Medicine, Surgery, and Public Health*, *2*, Article 100051. https://doi.org/10.1016/j.glmedi.2024.100051

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality*. John Wiley & Sons. https://doi.org/10.1002/0471458740

Blumer, H. (1954). What is wrong with social theory? American Sociological Review, 19(1), 3–10. https://doi.org/10.2307/2088165

Bosch, O. J., & Maslovskaya, O. (2023). *GenPopWeb2: The utility of probability-based online surveys - Literature review*. NCRM. https://www.ncrm.ac.uk/documents/GenPopWeb2_The%20utility%20of%20probability-based%20online%20surveys_Literature%20review.pdf

Bosnjak, M., Das, M., & Lynn, P. (2016). Methods for Probability-Based Online and Mixed-Mode Panels: Selected Recent Trends and Future Perspectives. *Social Science Computer Review, 34*(1), 3-7. https://doi.org/10.1177/0894439315579246

Bosnjak, M., Haas, I., Galesic, M., Kaczmirek, L., Bandilla, W., & Couper, M. P. (2013). Sample composition discrepancies in different stages of a probability-based online panel. *Field Methods, 25*(4), 339–360. https://doi.org/10.1177/1525822X12472951

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

Buys, T., Casteleijn, D., Heyns, T., & Untiedt, H. (2022). A reflexive lens on preparing and conducting semi-structured interviews with academic colleagues. *Qualitative Health Research*, *32*(13), 2030-2039.

Callegaro, M., Lozar Manfreda, K., & Vehovar, V. (2015). *Web survey methodology*. SAGE Publications Ltd. https://www.websm.org/

Chang, L., & Krosnick, J. A. (2009). National surveys via RDD telephone interviewing versus the Internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly, 73*(4), 641-678. https://doi.org/10.1093/poq/nfp075

Coast, J., Al-Janabi, H., Sutton, E. J., Horrocks, S. A., Vosper, A. J., Swancutt, D. R., & Flynn, T. N. (2012). Using qualitative methods for attribute development for discrete choice experiments: Issues and recommendations. *Health Economics, 21*(6), 730–741. https://doi.org/10.1002/hec.1739

Cornesse, C., & Blom, A. G. (2020). Response Quality in Nonprobability and Probability-based Online Panels. *Sociological Methods & Research*, *52*(2), 879-908. https://doi.org/10.1177/0049124120914940

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J. W., Struminskaya, B., & Wenz, A. (2020). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *Journal of Survey Statistics and Methodology*, *8*(1), 4–36. https://doi.org/10.1093/jssam/smz041

Daikeler, J., Bošnjak, M., & Lozar Manfreda, K. (2020). Web versus other survey modes: An updated and extended meta-analysis comparing response rates. *Journal of Survey Statistics and Methodology, 8*(3), 513–539. https://doi.org/10.1093/jssam/smz008

Eden, A. B., & Inan, N. G. (2022). Common Misconceptions and Misunderstandings in Magic Cut-Off for Significance: P-Value. In *Proceedings of the 4th International Conference on Statistics: Theory and Applications* (pp. 1-4).

Erfanian, F., Latifnejad Roudsari, R., Haidari, A., & Noghani Dokht Bahmani, M. (2020). A narrative on using vignettes: Its advantages and drawbacks. *Journal of Midwifery and Reproductive Health*, *8*(2), 2134–2145. https://doi.org/10.22038/jmrh.2020.41650.1472

Eurostat. (2021). *ESS handbook for quality and metadata reports: 2021 re-edition*. Publications Office of the European Union. https://doi.org/10.2785/965158

Granovetter, M. (1985). Economic Action and Social Structure: The Problem of Embeddedness. *American Journal of Sociology*, *91*(3), 481–510. http://www.jstor.org/stable/2780199

Groves, R. M. (1989). *Survey errors and survey costs*. John Wiley & Sons. https://doi.org/10.1002/0471725277

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). An introduction to survey methodology (Chapter 1). In *Survey methodology* (pp. 1-38). Wiley.

Hood, C. (2011). *The blame game: Spin, bureaucracy, and self-preservation in government*. Princeton University Press.

Ivanovska, A., Bosnjak, M., & Vehovar, V. (2025). *Relative bias in health estimates from probability-based online panels: Systematic review and meta-analysis*. Journal of Health and Rehabilitation Sciences, 4(1), 1–14. https://doi.org/10.33700/jhrs.4.1.137

Ivanovska, A., Bosnjak, M., & Vehovar, V. (2026). *Data quality in estimates from probability-based online panels: Systematic review and meta-analysis*. Acta Informatica Pragensia, 15(1), 173–197. https://doi.org/10.18267/j.aip.279869

Kocar, S., & Biddle, N. (2023). Do we have to mix modes in probability-based online panel research to obtain more accurate results? *Methods, Data, Analyses, 17*(1), Article 11. https://doi.org/10.12758/mda.2022.11

Lehdonvirta, V., Oksanen, A., Räsänen, P., & Blank, G. (2021). Social media, web, and panel surveys: using non-probability samples in social and policy research. Policy & internet, 13(1), 134-155.

Lincoln, Y. S., & Guba, E. G. (1986). *But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation*. New Directions for Program Evaluation, 1986(30), 73–84. https://doi.org/10.1002/ev.1427

Lobe, B., Morgan, D. L., & Hoffman, K. (2022). A Systematic Comparison of In-Person and Video-Based Online Interviewing. *International Journal of Qualitative Methods*, *21*. https://doi-org.nukweb.nuk.uni-lj.si/10.1177/16094069221127068

Lozar Manfreda, K., Bosnjak, M., Berzelak, J., Haas, I., & Vehovar, V. (2008). Web surveys versus other survey modes: A meta-analysis comparing response rates. *International Journal of Market Research, 50*(1), 79–104. https://doi.org/10.1177/147078530805000107

Lyberg, L. E., & Stukel, D. M. (2017). The roots and evolution of the total survey error concept. In P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, & B. T. West (Eds.), *Total survey error in practice* (pp. 3–18). Wiley.

MacInnis, B., Krosnick, J. A., Ho, A. S., & Cho, M.-J. (2018). The accuracy of measurements with probability and nonprobability survey samples: Replication and extension. *Public Opinion Quarterly, 82*(4), 707-744. https://doi.org/10.1093/poq/nfy038

Mercer, A., & Lau, A. (2023, September 7). *Comparing two types of online survey samples*. Pew Research Center. https://www.pewresearch.org/methods/2023/09/07/comparing-two-types-of-online-survey-samples/

Nickel, J., Hurst, A., & Duimering, P. R. (2024). Contextual influences on trade-offs in engineering design: a qualitative study. *Design Science*, *10*, e21. https://doi.org/10.1017/dsj.2024.34

Olson, K., Wagner, J., & Anderson, R. (2021). Survey costs: Where are we and what is the way forward? *Journal of Survey Statistics and Methodology, 9*(5), 921–942. https://doi.org/10.1093/jssam/smaa014

Pennay, D. W., Neiger, D., Lavrakas, P. J., & Borg, K. (2018). *The Online Panels Benchmarking Study: A total survey error comparison of findings from probability-based surveys and non-probability online panel surveys in Australia (CSRM Methods Series No. 2/2018).* Centre for Social Research and Methods.

Rohr, B., Felderer, B., Silber, H., Daikeler, J., Roßmann, J., & Schröder, J. (2024). *When are non-probability surveys fit for my purpose?* (Version 1.0). GESIS – Leibniz Institute for the Social Sciences. https://doi.org/10.15465/gesis-sg_en_050

Schwarz, G., Christensen, T., & Zhu, X. (2022). Bounded rationality, satisficing, artificial intelligence, and decision-making in public organizations: The contributions of Herbert Simon. *Public Administration Review*, *82*, 902.

Simon, H. A. (1955). *A behavioral model of rational choice*. The Quarterly Journal of Economics, 69(1), 99–118. https://www.jstor.org/stable/1884852

Simon, H. A. (1990). Bounded rationality. In *Utility and probability* (pp. 15-18). Palgrave Macmillan, London.

Tavory, I., & Timmermans, S. (2014). *Abductive analysis: Theorizing qualitative research*. University of Chicago Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Unangst, J. J., Amaya, A. E., Sanders, H. L. P. S., Howard-Doering, J., Ferrell, A. R., Karon, S. L., & Dever, J. A. (2020). A process for decomposing total survey error in probability and nonprobability surveys: A case study comparing health statistics in US internet panels. *Journal of Survey Statistics and Methodology, 8*(1), 62-88. https://doi.org/10.1093/jssam/smz040

Vehovar, V & Čehovin, G.: Izzivi uporabe neverjetnostih spletnih panelov v družboslovnem raziskovanju, Teorija in Praksa, 59, 4/2022, https://repozitorij.uni-lj.si/IzpisGradiva.php?id=144123

Vehovar, V., & Beullens, K. (2018). *Cross-national issues in response rates*. In D. L. Vannette & J. A. Krosnick (Eds.), The Palgrave handbook of survey research (pp. 29–42). Palgrave Macmillan. https://doi.org/10.1007/978-3-319-54395-6_6

Vehovar, V., Berzelak, N., & Lozar Manfreda, K. (2010). *Mobile phones in an environment of competing survey modes: Applying metric for evaluation of costs and errors*. Social Science Computer Review, 28(3), 303–318. https://doi.org/10.1177/0894439309350634

Vehovar, V., Toepoel, V., & Steinmetz, S. (2016). *Non-probability sampling* (Vol. 1, pp. 329-345). The Sage handbook of survey methods.