Public Administration Review

PAR Public Administration Review ASPA

RESEARCH ARTICLE OPEN ACCESS

The Effect of Performance Failures on User Satisfaction: Evidence From a Natural Experiment

Mads Thau[1] | Maria Falk Mikkelsen[2] | Nathan Favero[3]

[1]Institute for Social Research, Oslo, Norway | [2]VIVE – The Danish Center for Social Science Research, Copenhagen K, Denmark | [3]American University, Washington, D.C., USA

Correspondence: Mads Thau (mads.thau@samfunnsforskning.no)

Received: 18 October 2024 | **Revised:** 17 October 2025 | **Accepted:** 22 October 2025

Funding: This work was supported by the National Board of Social Services in Denmark.

Keywords: natural experiment | performance failure | satisfaction

ABSTRACT

Despite long-standing interest in satisfaction with public services and organizations, our knowledge of how responsive user satisfaction is to real-world performance fluctuations remains limited. Existing cross-sectional studies may suffer from selection bias, while survey experiments may overstate performance information effects, as the salience of such information is artificially primed. We exploit a unique opportunity to study the link between performance failure and user satisfaction dynamically, as news of a major performance failure within the Danish National Board of Social Services happened to break during fielding of a satisfaction survey among the Board's users. Our analysis shows no negative effects of the performance failure on user satisfaction. These findings suggest that in real-world settings—where citizens draw on many information sources when forming judgments—performance effects on satisfaction are weaker than prior studies suggest. Thus, satisfaction data cannot be assumed to automatically reflect changes in service providers' performance and reputation.

Satisfaction surveys among users are widely used to measure the subjective aspects of service quality and organizational performance in the public sector (Olsen 2015; Grøn and Kristiansen 2022; Song et al. 2025). Their value partly rests on the assumption that there is a close relationship between performance and satisfaction (Stipak 1979). Take a researcher asked to complete a survey evaluating the application and disbursement process of their national research funding organization as an example. If, shortly before responding, this researcher learned that a failure in the organization's internal control systems had allowed an employee to embezzle millions of the funds that were ear-marked for legitimate research, the prevailing view in the literature is that the researcher would report low satisfaction to signal discontent. Indeed, such a translation of poor performance into low satisfaction ratings provides essential feedback for political and administrative principals, as it may be the primary source of performance information available to support their oversight of public service delivery.

Theoretically, the expectancy-disconfirmation model (EDM)—the predominant approach to explain satisfaction with public services—affirms the importance of (perceived) performance for user satisfaction, but also posits that users (e.g., citizens, clients, companies, charities) compare the performance of a service against their expectations of that service. High satisfaction is predicted if the perceived performance meets or exceeds expectations; low satisfaction follows when expectations are disconfirmed (Van Ryzin 2013; Hjortskov 2017). Numerous empirical studies have examined the relevance of the EDM for understanding the link between performance and user satisfaction either with cross-sectional data on real user satisfaction or

Evidence for Practice

- Be cautious. Like any performance metric, quantitative user satisfaction data gives at best a partial view of things.
- Don't assume. Users can interpret satisfaction surveys in many ways and can take them as an opportunity to voice any of a variety of opinions they have.
- Be proactive. Consider including open-ended questions in user surveys or following up with qualitative interviews for a broader perspective on what users are experiencing.

by experimentally manipulating variables believed to influence satisfaction (see Zhang et al. 2022).

However, while these studies have provided important evidence on the usefulness of the EDM in public administration research, the available evidence has limitations. First, the available cross-sectional studies do not allow for precise causal conclusions to be drawn due to the potential for omitted variable and selection biases (e.g., since users often have some agency in choosing providers). Second, as the effects of performance on user satisfaction are found to be stronger when respondents are primed (Andersen and Hjortskov 2016), experimental studies that provide obvious performance cues may overestimate the importance of performance information. For example, Marvel (2016) has shown that performance information cues impact respondents' performance perceptions, but these effects are extremely short-lived. Thus, while survey experiments have many advantages, there are external validity concerns regarding how performance effects play out in the real world, with all its complexity and informational richness.

This article leverages a natural experiment to consider how real-world performance failures affect user satisfaction. While the EDM suggests that performance failure (performance well below expectations) will lower user satisfaction, we also find in the literature alternative perspectives that suggest the performance-satisfaction link may not be as straightforward as typically assumed. Several studies document cognitive biases in how users respond to performance information, such as the use of motivated reasoning, whereby individuals are reluctant to accept information that challenges their pre-existing judgments because it is uncomfortable to change beliefs about the world (Baekgaard and Serritzlew 2016; Jilke and Tummers 2018; see also Stipak 1979). Another perspective argues that people rationally update their beliefs only incrementally based on a single new piece of information, drawing on a logic of Bayesian updating amid informational uncertainty (Hjortskov 2019). Under this view, adjustments to satisfaction should be particularly muted when individuals have firmly held pre-existing beliefs (strong priors), for example after accumulating many observations of an organization's performance. Thus, there are theoretical grounds to question whether a single performance failure will consistently have notable effects on real-world satisfaction.

Our empirical study examines a large performance failure in Denmark, involving a government agency called the National Board of Social Services (NBSS). Users served by the agency are non-profit charities, social enterprises, and local government organizations, which the NBSS is contractually obliged to survey in order to report on user satisfaction to its overseeing ministry (Jensen et al. 2021). To test the effect of the performance failure, we use data from a two-round satisfaction survey carried out among NBSS users in 2018 and 2020. About halfway through the 2018 data collection, the media revealed a scandal involving a senior NBSS employee who had embezzled DKK 117 million (~USD 18 million) over a span of 25 years. Since NBSS users rely on funding from the NBSS, they were directly affected by this performance failure, as the embezzled money came from transfers meant to support their projects. This long-standing embezzlement highlighted a systematic failure in the agency's control mechanisms and cast serious doubts on the integrity of funding decisions and payments experienced by users for decades. Unsurprisingly, the NBSS scandal led to massive media coverage, which portrayed the NBSS as an extreme case of what Andrews et al. (2007) call organizational "mismanagement".

We estimate the effect of the performance failure based on a pre-analysis plan inspired by recommendations from Muñoz et al. (2020) about how to study unexpected events occurring during survey collection. Specifically, we first use the 2018 survey to compare the mean satisfaction of respondents who completed the survey before and after the performance failure became publicly known. We then implement a more sophisticated design where we use results from the 2020 round of the survey to control for systematic and unobserved differences between late and early replying respondents.

Our findings show no consistent evidence of negative effects of the performance failure, even when conducting subgroup analyses to probe for heterogeneous effects. We also show that these null effects are unlikely to result from insufficient statistical power. Any effects our study might be unable to detect would be substantively small and unimportant. We end the article by discussing the theoretical and practical implications of these findings.

1 | Service Users and Performance Information

Service users—those individuals or organizations receiving a particular public service—are in many cases the key stakeholders in the public administrative sphere. Service users rarely have formal power vis-à-vis service providers. Unlike voters, they usually cannot 'throw out the rascals' or otherwise directly sanction professionals, bureaucrats, or agencies for performance failures. They may not even be able to 'vote with their feet' and choose another provider due to monopolies on particular services (e.g., decisions on transfers, permits, grants, sanctions, and so forth).

Service users nonetheless have other powers in that they play a key role in informing principals about an agent's performance by expressing their views on the public services received from the agent (Damgaard and Nielsen 2020; Moynihan 2008). Such performance information is crucial in reducing the information asymmetries within governmental hierarchies (Brehm and Gates 1997), but it is also relevant in keeping governments accountable to the public more generally (Djerf-Pierre et al. 2013).

User satisfaction surveys have become a standard tool for measuring subjective aspects of performance (Olsen 2015; Van de Walle and Van Ryzin 2011), with wide use by researchers and practitioners alike in settings such as schools, hospitals, local governments, or agencies. For example, alongside other sources of information, user satisfaction is central to modern performance management systems because user assessments are believed to meaningfully reflect good and bad performance (Song and Meier 2018; Schachter 2010). At the very least, satisfaction surveys offer decision-makers unique insights into how services actually work at the receiving end (Hjortskov 2017; Kelly 2005; Stipak 1979), helping organizations improve future service provision. From a normative standpoint, satisfaction surveys give citizens the opportunity to 'voice' their specific wishes and concerns to service providers and political superiors, supplementing the crude signals they are otherwise able to send as voters every four or five years (Hirschman 1970; Mulgan 2000). Given these considerations, it is unsurprising that recent studies find that top managers in public organizations pay a good deal of attention to input from user satisfaction surveys when making decisions (Grøn and Kristiansen 2022; Moynihan and Hawes 2012).

2 | The Dominant View: The Expectancy-Disconfirmation Model

The service feedback mechanism hinges on users' ability to respond to changes in performance standards and express dissatisfaction if performance falls below expectations. There is both observational and survey experimental research to suggest that performance below user expectations is generally linked to lower satisfaction (Van Ryzin 2013; Zhang et al. 2022).

More broadly, the EDM identifies three key variables as drivers of satisfaction: (perceived) performance, expectations, and 'disconfirmation'—a comparison of the performance level with expectations (Oliver 1980). Under the model, performance and expectations can each affect satisfaction both directly and indirectly, with indirect effects being mediated by disconfirmation. Intuitive interpretations of direct versus indirect effects typically equate direct performance effects on satisfaction as reflecting how *absolute* levels of performance affect satisfaction, whereas indirect performance effects mediated by disconfirmation represent satisfaction's responsiveness to performance in a *relative* sense (the performance level relative to expectations).

Thus, under the EDM, performance is supposed to be positively related to satisfaction, clearly suggesting the following hypothesis about performance failure:

Hypothesis 1. *Mean satisfaction is lower among users exposed to a performance failure than among users not exposed to it.*

3 | Heterogeneity in Responses: The Role of Experience

The above hypothesis describes an average effect, but there may be considerable heterogeneity in how individuals form satisfaction with a public service provider. One particularly obvious factor to consider is how much experience an individual has interacting with a service provider, since a greater level of interaction implies more first-hand experience that one can draw on when evaluating an organization. Indeed, one justification for the importance of expectations to satisfaction in the EDM directly invokes the example of users with little experience to draw upon. Van Ryzin (2006) argues that if citizens lack meaningful information about performance, they may fall back on 'imputing' missing performance data with their expectations for performance, leading to a positive direct effect of expectations on satisfaction (sometimes called an assimilation effect). One might argue that the sudden unveiling of vivid information about performance (as in the case of a widely publicized scandal) should have a particularly strong effect for individuals who lack much first-hand knowledge of a service provider. Meanwhile, individuals with substantial prior experience with the provider may find less need to update their bottom-line satisfaction judgments when new information emerges, seeing a scandal as a potential one-off example.

Another possibility, however, is that those with high levels of experience with an organization may feel the effects of a scandal most vividly. Being highly experienced, they may be particularly attentive to the ways in which they were affected firsthand or could have easily been affected under slightly different circumstances. While the potential heterogeneity of responses to performance failure is not the main focus of this article, our empirical analysis will include supplemental analyses testing for heterogeneous effects by user experience levels (in line with our pre-analysis plan).[1]

4 | A Competing View? Emerging Criticisms of the EDM

Despite the ubiquity of the EDM in satisfaction research, some studies provide grounds for questioning its dominance. Perhaps the most direct critique is that of Favero and Kim (2021), who criticize the model's treatment of disconfirmation as a unique construct, arguing that it is neither empirically nor conceptually distinct from other constructs in the model (specifically expectations and performance). More recently, Favero et al.'s (2025) analysis of panel data on satisfaction indicates that the link between expectations and satisfaction is rather weak and short-lived, a finding at odds with EDM's characterization of expectations as a central driver of satisfaction.

Another critique comes from Andersen and Hjortskov (2016), who argue that findings of cognitive biases in performance evaluations undermine the EDM, since it is a model based on the assumption of a rational cognitive process. They offer as an alternative a dual-process model that can accommodate intuitive judgments and not just deliberative, rational processing of information (see also Hjortskov 2017, 2019).

One particularly notable source of bias is motivated reasoning, which is a tendency to accept or interpret information in ways that conform to pre-existing beliefs (Baekgaard and Serritzlew 2016; Jilke and Tummers 2018), implying that attitudes may be difficult to meaningfully change with new information. An alternative perspective is a Bayesian model of satisfaction judgments

(Hjortskov 2019; Favero et al. 2025). Under this model, when encountering new information, service users incorporate it as new data that can help to update their Bayesian "prior" beliefs about possibilities for quality/performance in the absence of complete information. In this way, a user's satisfaction reflects a cumulative judgment regarding their past experiences with a service provider, and views may eventually become so entrenched that there is little reaction to a single instance of new performance information.

When it comes to empirical evidence backing the EDM's claim of a reliable link between performance and satisfaction, early studies of satisfaction initially cast doubt on the responsiveness of citizen satisfaction to performance, finding weak or insignificant links between the two variables (Brown and Coulter 1983; Stipak 1979; Kelly 2003). More recent work, however, appears to generally support the notion of a strong relationship between performance and satisfaction (Zhang et al. 2022). Yet most studies with strong causal research designs employ survey experiments based on fictional vignettes that may provide respondents with little information other than performance cues on which to base their opinions. Absent other meaningful information, it is hardly surprising that performance drives satisfaction. Although a few experimental studies take a more real-world approach by estimating the effects of disclosure of true performance information about actual organizations (James and Moseley 2014; Damgaard and Nielsen 2020), these studies may also artificially inflate reactions to performance information because respondents are primed to think about such information shortly before they assess performance.

Given the above concerns about both empirical evidence and theoretical underpinnings, one might reasonably question the EDM's dominance as an organizing framework for studying user satisfaction in public administration. While this article does not offer any formal hypotheses against the EDM, we take seriously the possibility of a null effect as an alternative to H1. Despite the dominance of the EDM, the various linkages in the model have not been consistently supported across different types of research designs.

5 | Case: The NBSS and the 2018 Embezzlement Case

To study the effect of performance failure on user satisfaction, we exploit the fact that news of a major embezzlement scandal in the NBSS unexpectedly broke during the data collection stage of a biannual satisfaction survey among NBSS users.

The NBSS is a large government agency responsible for a variety of tasks in the social policy area, including funding and oversight of social work projects. The NBSS receives and processes between 1500 and 1800 funding applications yearly from user organizations, mainly non-profit charities, voluntary organizations, social enterprises, or public organizations at the local governmental level. The NBSS provides users with information about funding opportunities, support during the application process, decisions and notifications on applications, disbursements of grant money, and financial oversight after grant payouts.

NBSS users can be categorized as collective actors with high levels of proficiency and resources (Jensen et al. 2021). Many of the employees at these organizations have master's degrees and extensive experience applying for NBSS funding. Because of limited opportunities for alternative funding, a negative decision from the NBSS could result in staff layoffs within user organizations. Drawing on Hirschman (1970), it is therefore reasonable to anticipate that NBSS users would utilize available avenues, such as anonymous satisfaction surveys, to voice their concerns and experiences given the limited exit opportunities.

5.1 | The Embezzlement Scandal

On October 9, 2018, news broke of an embezzlement scandal involving a senior employee at the NBSS. The employee had embezzled DKK 117 million (around EUR 16 million or USD 18 million, at the time) in grant funds—money that would otherwise have covered new projects or existing operating costs for NBSS users. The fraud took multiple forms, such as creating fictitious projects to divert funds, awarding applicants smaller grants than requested and redirecting the surplus, and imposing burdensome administrative requirements in order to reallocate unclaimed funds. In some cases, funds were stolen directly from legitimate projects (Ditzel 2021). As a result, many NBSS users were likely affected either directly—through stolen funds, reduced grant allocations, or bureaucratic hurdles that caused them to forgo funding—or indirectly, by facing artificial competition from fictitious projects.

News of the embezzlement scandal shocked the Danish public. Denmark is widely regarded as one of the least corrupt countries globally (Transparency International 2020), and a fraud of this size and longevity was previously unseen in a Scandinavian context. The embezzlement also directly affected projects intended to aid society's most vulnerable groups, such as homeless people, at-risk children, and battered women. As such, the "Britta Nielsen Scandal" (so named after the implicated NBSS employee) became one of the biggest Danish news stories of 2018 and constituted a powerful performance signal for NBSS users.

Figure 1 illustrates public interest in the performance failure at NBSS during the three-week period that corresponds to NBSS's 2018 satisfaction survey. At the macro level, the agency attracted mass media attention (operationalized as newspaper articles) throughout the entire post-performance failure period covered by our survey data, particularly in the first week (panel A). At the individual level, citizens were also occupied with the embezzlement, as indicated by the explosion in Google searches for the NBSS by Danish internet users after October 9, 2018 (panel B).

The narrative around the embezzlement was that it was as much a collective system failure as a personal criminal case, clearly attributed to "mismanagement" rather than "misfortune" (Andrews et al. 2007, 275). The embezzlement had been ongoing for at least 25 years, exposing a severe lack of procedural oversight and control mechanisms. The offending employee—a highly trusted member of staff—had super-user access to the grant application system and was able to disburse grant funds, act as auditor, and edit information in the system such as account numbers (Gottschalk 2021). During the press conference where

the embezzlement was publicly announced, the then-Minister of the Interior openly said that "internal control mechanisms had been failing for years to an unprecedented degree" and spoke of a pressing need to "clean up and tighten" NBSS procedures. Indeed, the public, including the NBSS users, essentially saw the agency being put under administration by the Ministry of the Interior after the press conference.

More broadly, the NBSS was clearly held accountable for the performance failure both in the media and by its principal, leaving it little opportunity for blame avoidance and reputation management in the short run (Ditzel 2021). There were also examples of NBSS users directly linking the embezzlement scandal to their own experiences. For instance, a non-profit organization suggested on national news that the fraud could be responsible for its lack of funding and subsequent layoff of personnel (DR 2018). In an open-ended comment field at the end of the 2018 survey, we saw examples of respondents noting that their responses were influenced by the scandal.

6 | Data, Measures, and Analytical Strategy

Like many other agencies, the NBSS is required to collect data on user satisfaction by its overseeing principal (the Ministry of the Interior). We use a two-round user satisfaction survey among NBSS users conducted by the Danish Center for Social Science Research. The survey was sent to each organization's registered contact person at the NBSS. Recipients were encouraged to respond on behalf of their organization, forward the survey to a more appropriate colleague if necessary, or complete it collaboratively. Respondents were made aware that the survey was anonymous and thus had no reason to fear that replies would be connected to future funding decisions by the NBSS.

The first round was fielded October 1–22, 2018—covering the public announcement of the performance failure on October 9—while the second round ran exactly two years later. Although all respondents were invited to participate on October 1, many completed the survey at a later date. The survey aimed to track the development in overall user satisfaction as well as ratings for more specific services. Our main source of data is the first survey round, which was fielded among all user organizations that received funding between January 1, 2017 and July 15, 2018.

The second round, which allows us to control for potential differences between early- and late-survey respondents (see below), was fielded on October 1–22, 2020 among organizations receiving funding between January 1, 2019, and July 15, 2020. The response rates were 52% in 2018 ($n = 460$) and 55% in 2020 ($n = 472$), resulting in a pooled sample of 932 observations. In a test of representativeness, we find substantively small and statistically insignificant differences between respondents and non-respondents on available background variables (Supporting Information (SI), Section A).

We note that no records of unsuccessful applications are kept. Thus, the NBSS satisfaction survey was distributed to all user organizations that succeeded in securing at least some funding during the target timeframe. While most respondents have a long history of applying for funding, experiencing both successful and unsuccessful application attempts, our sample does not include users who only had unsuccessful application attempts. Consequently, one limitation to the external validity of our study is that we cannot infer what the performance failure meant for applicants that never received funding.

6.1 | User Satisfaction Measures

The survey measures user satisfaction based on the American Customer Satisfaction Index (Fornell et al. 1996), which has informed satisfaction surveys in existing public administration research (e.g., Jilke and Baekgaard 2020). Specifically, respondents were asked: "Please consider your full experience in terms of applying for, receiving, and administering public grants in the social area. Overall, how satisfied are you with the grant management of the National Board of Social Services?" Responses were then given on a 10-point scale from 1 "very dissatisfied" to 10 "very satisfied". The sample mean is 6.5 (SD = 2.2). We have

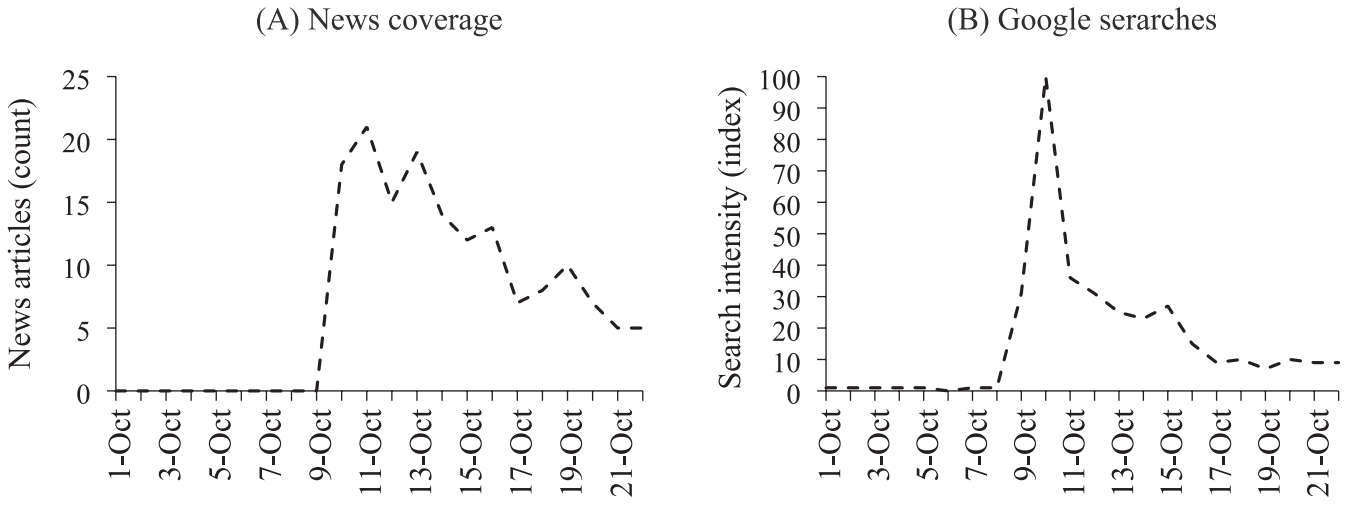

FIGURE 1 | Public interest in the NBSS around the embezzlement case in October 2018. Panel A shows the number of articles featuring the NBSS in national print media (Infomedia), and panel B shows unique searches for NBSS on Google, indexed from 0 to 100 with October 10, 2018 at 100 (Google).

rescaled satisfaction to range from 0 to 100 in the analysis so that estimates can be interpreted in percentage points (M = 61.4; SD = 24.5) but we also discuss standardized effects. The distributions on satisfaction before and after the scandal are shown in the SI (Section A).

6.2 | Analytical Strategy

Following calls for increased transparency in non-experimental research (Nosek et al. 2018), we conduct our analysis according to a pre-analysis plan available at the OSF site.[2] We estimate the effect of the embezzlement scandal on user satisfaction using a design—inspired by Muñoz et al. (2020)—which exploits that respondents in the 2018 survey were unaware of the scandal for the first eight days of the data collection leading up to the public announcement on October 9. For the last 14 days of the collection period, respondents had been exposed to the scandal.

Before estimating any models, we examine the satisfaction data descriptively. Figure 2 shows daily mean user satisfaction over the collection period in 2018 (square dots), as well as the trends in satisfaction before and after October 9 using local polynomial smooths (solid and dashed lines). It is difficult to identify any clear time trends, and even harder to see any sign of satisfaction dropping after the October 9 cutoff.[3] Still, effects may be suppressed by confounders and could therefore emerge once we estimate our full models.

Our first set of models uses OLS regression with robust standard errors to estimate the effect of the performance failure by simply comparing user satisfaction before and after October 9 using the 2018 data, thus treating the scandal as an exogenous shock to mean satisfaction. Control variables are included to account for the fact that the respondents choosing to participate before October 9 are not necessarily identical to those participating thereafter in terms of factors related to satisfaction. We include both individual-level factors (age, gender, education, and position) and organizational characteristics (type of organization, type of grant received, historical experience with NBSS services, and self-reported competencies in the organization related to applying for and managing grants). We specify the model on the full 2018 sample but also on a restricted sample using only responses given right before (October 5–8) and right after (October 9–15) the embezzlement scandal.[4]

As our second and main (pre-registered) approach, we use data from both 2018 and 2020 to estimate the following model with robust standard errors:

$$Y_{it} = \alpha + \beta_1 \cdot Year_t + \beta_2 \cdot After_i + \delta \cdot \left(After_i \cdot Year_t\right) \cdot + X'_{it} \cdot \gamma + \varepsilon_{it}$$

Here Y_{it} is the level of user satisfaction. $Year_t$ is a dummy variable equal to 1 for observations from 2018, and $After_i$ is a dummy variable measuring whether respondents answered the survey before or after October 9 (indicating exposure to the performance failure). $After_i \cdot Year_t$ is the interaction term between the two variables, with δ being the coefficient of interest capturing the effect of the performance failure.

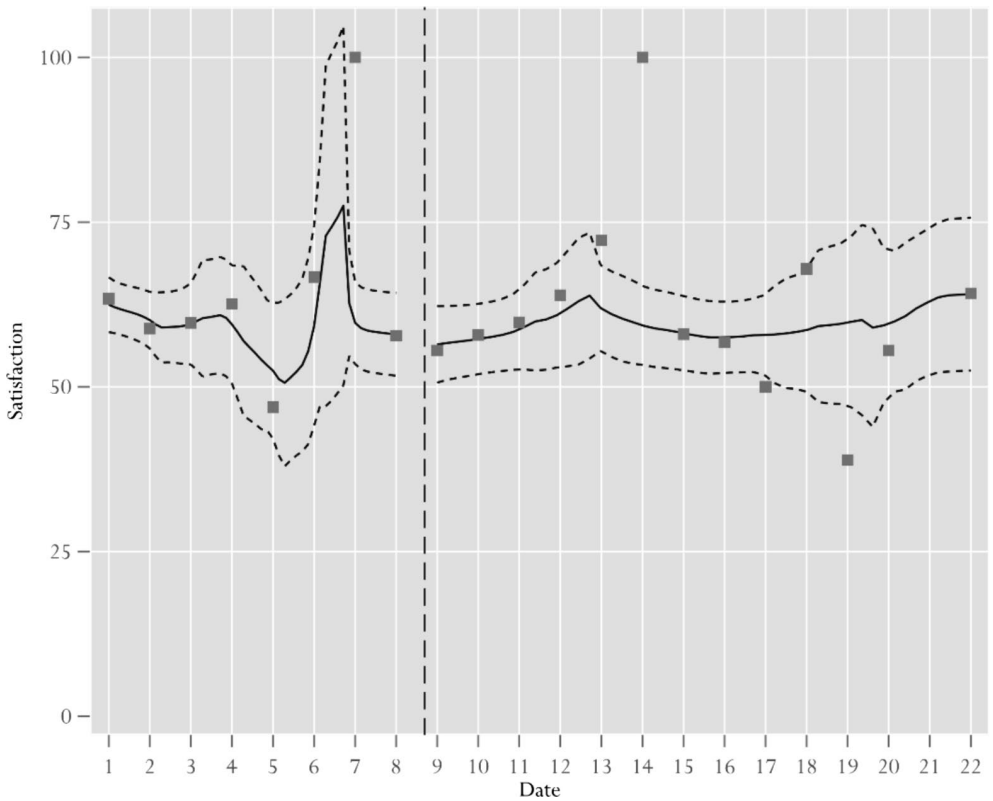

FIGURE 2 | Satisfaction among surveyed NBSS users in October 2018. Squares are daily means. Solid lines are local polynomial smooths and dashed lines their 95% confidence intervals. These are based on, respectively, the period before (left) and after (right) the NBBS scandal. Note that the means on October 7 and 14 represent a single response.

With this model, we estimate the difference in user satisfaction before and after October 9 in 2018 *net* of the difference in user satisfaction before and after October 9, 2020 (i.e., the counterfactual trend). Thus, we control for general differences between early and late replying respondents occurring in both years (before/after fixed effects), and for any general differences between the 2018 and 2020 samples (survey year fixed effects), alongside the included covariate vector X'_{it} (same variables as above).

The fact that the embezzlement scandal was known to everyone in the 2020 sample does not bias our estimates, even if the scandal led to permanently lower user satisfaction in absolute values in 2020, as such year differences are canceled out by design. We nevertheless note that the NBSS received much less attention from both the media and citizens during the data collection period in 2020 compared to 2018, suggesting that there was little spill-over from 2018 to 2020 (see SI, Section B).

A distinct concern is whether the performance failure influenced respondents' likelihood of completing the survey. On the one hand, the failure might have increased participation by motivating users to voice complaints; on the other, it could have discouraged participation due to frustration or disillusionment. As mentioned above, we tested for the representativeness of our sample overall. But to address the present concern, we also examined whether respondent characteristics shifted after October 9, 2018, benchmarking against trends from the 2020 survey. If the performance failure affected survey participation, we would expect to observe shifts in respondent characteristics in 2018 that are different from 2020. We find no such (significant) differences and thus no evidence that news of the performance failure altered who responded to the survey (SI, Section A).

Per our pre-analysis plan, we estimate models with and without covariates. The inclusion of control variables accounts for dynamic factors that may vary between the 2018 and 2020 samples. These are not picked up by the year fixed effects and pose a risk of bias if they co-vary with the performance failure indicator. The SI provides overviews and descriptive statistics of all variables in our models, as well as a discussion of the key identifying assumptions in our analysis (including balance tests, common trends tests, and placebo tests) and a power analysis.

7 | Findings

Did the NBBS performance failure lower user satisfaction? Using the 2018 survey, Figure 3 shows no significant effect of the performance failure, neither when we examine all data from 2018 (model 1) nor when the time window is reduced to right around the scandal. The estimates are even positive and thus in the opposite direction of expectations.[5]

Such a before-and-after comparison reflects the true effect of the performance failure, if—in absence of the performance failure—mean satisfaction for the late and exposed repliers would have been exactly equal to the mean satisfaction of early and unexposed repliers. Early- and late-repliers may, however, differ on several important factors that may affect user satisfaction.[6] Using the 2020 survey, we can estimate the general difference in satisfaction between early- and late-repliers, and then see

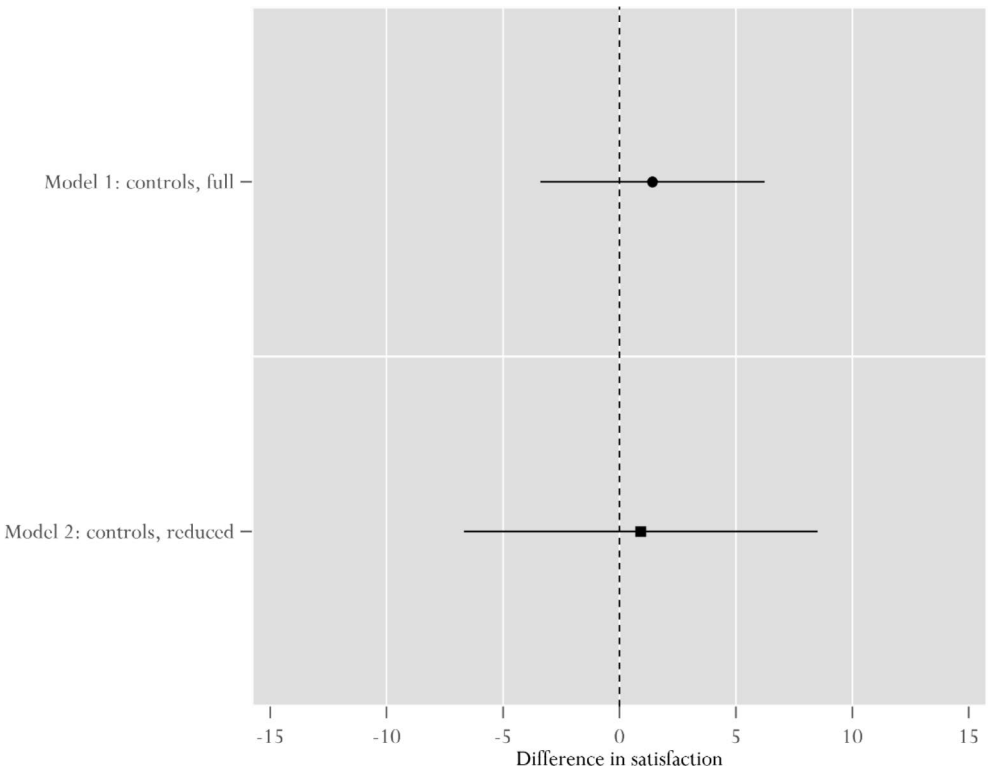

FIGURE 3 | The difference in overall satisfaction before and after October 9 in 2018. The dot and square are estimates and bands are 95% confidence intervals. Model 1 uses the full 2018-sample, while model 2 uses the reduced sample to zoom in around the scandal (before: Oct 5–8; after: Oct 9–15). See SI, Section B for full results.

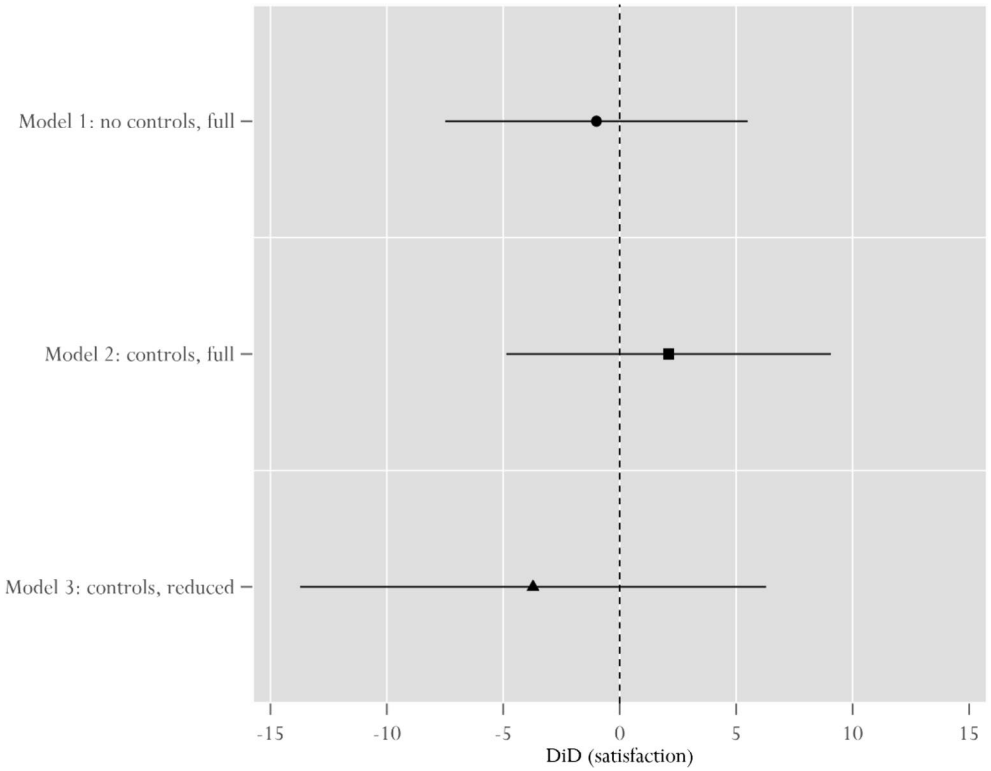

FIGURE 4 | The effect of the NBSS embezzlement case on overall satisfaction. The dot, square, and tringle are difference-in-differences estimates and bands are 95% confidence intervals. Model 1 uses the full before and after periods. Model 2 also uses the full before and after periods but includes control variables. Model 3 uses the reduced before and after periods and includes controls. See SI, Section B for full results.

whether there is a drop in satisfaction in 2018 relative to this baseline difference in early versus late survey responses.

Figure 4 reports our estimates using models that control for these potential differences between early and late survey responders (for full results, see SI, Section B). Consistent with the pre-analysis plan, we consider these more sophisticated models—specifically models 1 and 2—to be our primary tests. Overall, it is hard to see any consistent evidence that the performance failure of the NBSS did anything to lower user evaluations. In model 1, the point estimate of the effect of the scandal is one percentage point and highly uncertain ($\beta = -1.00$; $p = 0.76$). While we cannot rule out the possibility of a minor dip in satisfaction, the 95% confidence interval suggests that it is unlikely that such a drop in satisfaction would exceed 7.5 percentage points, or 0.31 standard deviations.

Model 2 includes our control variables, thus accounting for the possibility that the respondent composition across the before- and after-periods in 2018 was not completely identical to the composition in 2020. However, in model 2, a significant effect is still absent ($\beta = 2.10$; $p = 0.55$). The confidence interval for this model suggests an even smaller range of negative values, ruling out any negative effect of 5 percentage points (0.2 standard deviations) or more as unlikely. Our preregistered models thus contradict H1.

In model 3, we use the narrower time window around the scandal. This gives a negative but insignificant estimate ($\beta = -3.73$; $p = 0.46$) with a lower bound on the confidence interval of around 14 percentage points. Thus, although this exploratory model makes it more difficult to rule out negative effects of notable size (e.g., 10 percentage points), there is certainly no convergence on a clear negative effect.

Following our pre-analysis plan, we also considered heterogeneous effects. Even if many respondents were unaffected by the NBSS failure, certain subgroups may have expressed lower satisfaction ratings. In Figure 5, we focus on experience and parse out estimates by whether users have less or more than 10 years of experience with the NBSS. In the SI, we also check for heterogeneity by users' competences, the respondent's education, and the survey's question order (Section B). We find that none of the investigated subgroups signaled discontent with the NBSS by significantly lowering satisfaction ratings following the news of the scandal. Thus, the main result of null effects stands, even after accounting for experience and other factors as potential moderators.

8 | Additional Considerations: Robustness and Power

To test the robustness of our models, we explored the effects of the performance failure on alternative outcomes. The models in Figure 4 were rerun using a formative index based on respondents' ratings of six specific services provided by the NBSS.[7] The effect of the scandal remains insignificant and inconsistent whether we consider the satisfaction index or each of the rating questions separately (see SI, Section B for results and distributions on the rating questions).

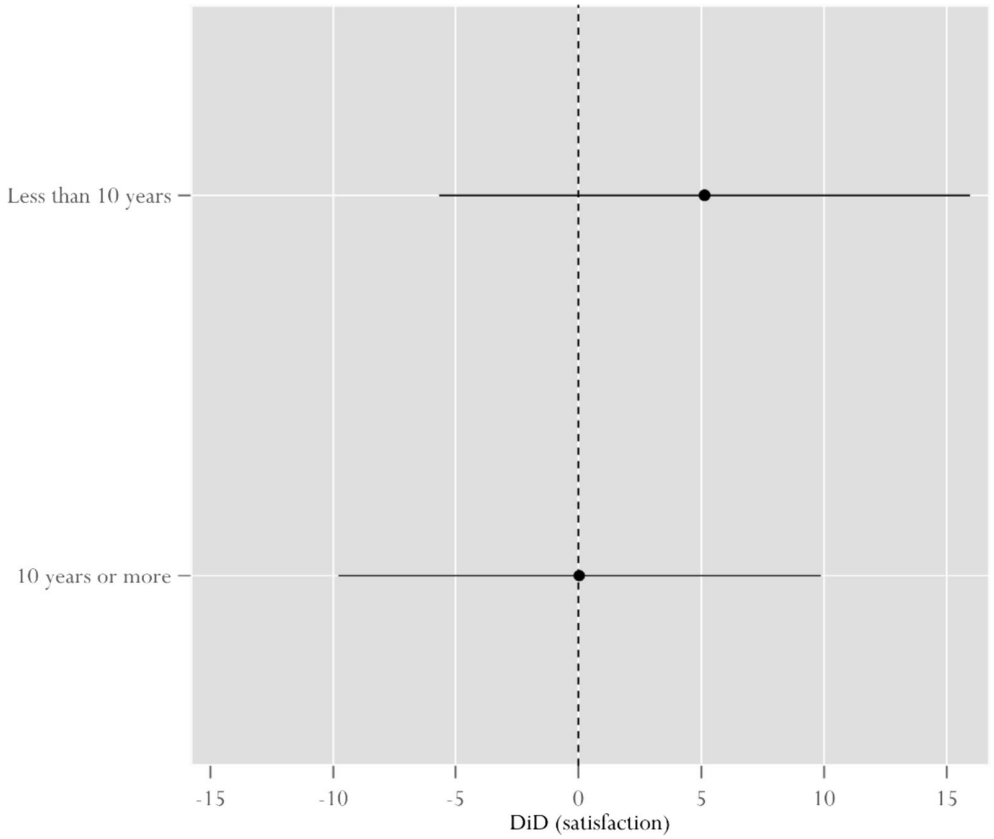

FIGURE 5 | The effect of the NBSS embezzlement case among more and less experienced user organizations. Dots are estimates and bands are 95% confidence intervals. The model specification is identical to Model 2 in Figure 4 (controls, full sample), except the inclusion of the higher-order interaction with experience (and constitutive terms). The difference in effect between more or less experienced users is insignificant ($p = 0.50$). $N = 743$. See SI, Section B for full results.

We also assessed whether the analysis might be underpowered. A sample of 932 users is by no standards small, and, in practice, such a sample size is what many satisfaction surveys would have to uncover effects (such as those we are able to detect for several covariates). We conducted formal power analyses, benchmarking off prior satisfaction studies and a common rule of thumb for effect sizes in the social sciences (see SI, Section A). Among studies documenting significant biases in performance assessments, effects range from 0.17 to 0.83 standard deviations, with most close to 0.2. For an effect size of 0.2 standard deviations, our power is estimated at 68%. Thus, our study is likely able to detect effects of the size typical for survey factors known to bias results, but which ideally would be irrelevant to rational satisfaction judgments.

9 | Discussion

Despite the evident performance failure of NBSS, we found no negative effects on users' satisfaction. While a true performance failure effect could in principle exist but simply be too small for our study to pick up, such an effect would be of minor substantive interest, as our sample size allows us to identify even small effects. From our 95% confidence intervals (with our main models), it appears that satisfaction could have dropped at most by 0.2 or 0.3 standard deviations in response to this widely publicized failure. Thus, any possible effect of the new substantive performance information being revealed in our empirical case can be reasonably interpreted as no larger than the typical noise from cognitive biases linked to the particular context or framing of a survey instrument measuring satisfaction (typically around 0.2 standard deviations; e.g., Van de Walle and Van Ryzin 2011; Hjortskov 2017; Thau et al. 2021). Furthermore, even if subtle effects of the performance failure existed in this case, they would remain unnoticed in practice, as regular-sized satisfaction surveys ($N \sim 1000$) would lack the power to detect them.

Our case is arguably more favorable for finding performance effects than most. The scandal received widespread national media coverage and directly affected users, many of whom plausibly lost funding opportunities as a result. Moreover, NBSS faced clear public condemnation for allowing the fraud to go undetected for over two decades. If such a high-profile, large-scale failure did not meaningfully reduce user satisfaction, then we must ask under which conditions a performance failure would affect user satisfaction in the real world. If the main way to detect the expected performance effects is to explicitly prime respondents with unambiguous information about a performance failure (as in survey experiments), we must question whether this approach truly enhances our understanding of how satisfaction judgments are formed by real service users and citizens.

We believe our findings lead to three key insights. First, our findings show that user satisfaction is more sluggish than envisioned by the EDM framework. In a real-world setting like the one we observe, users draw on a broader and more diffuse

set of experiences than is typically considered in survey experiments, including personal interactions with the organization, pre-existing beliefs, and background knowledge. New information is filtered through these prior experiences, resulting in incremental rather than dramatic shifts in perception. This could align with a Bayesian updating logic under conditions where individuals hold strong priors, such that new evidence marginally modifies (at undetectable levels), rather than overturns, existing views (Hjortskov 2019). Moreover, motivated reasoning may further blunt the impact of new information, as individuals often resist revising their beliefs in ways that induce psychological discomfort (Kunda 1990). Support for this view comes from a recent survey experimental study by Damgaard and Nielsen (2020, 9), who find that the satisfaction of experienced users (school parents) is unresponsive to performance cues. But they also argue that responsiveness might emerge if the "performance failure signals" were "more clear and consequential" than a dip in students' average grades. Our study shows that even in the presence of a highly publicized, large-scale organizational failure, user satisfaction remains largely unaffected. The fact that both studies—one survey experimental and one utilizing a natural experiment—find that experienced users are unresponsive to performance failures (in contrast to James and Moseley's 2014 finding among inexperienced users) suggests that experience levels may play a critical role in shaping satisfaction. While we explicitly tested for differences between relatively more and less experienced users in our case and found no evidence of heterogeneous effects, our sample's variation is somewhat constrained since all respondents had successfully secured funding, indicating a minimum level of competence and experience. Future research should examine settings where lower levels of user experience are observed to better understand potential heterogeneity among respondents, thus shedding light on important mechanisms underlying citizen and user evaluations. While our findings do not invalidate the EDM framework, they suggest that it may be limited as a lens for understanding how experienced users respond to performance failures in the real world.

Second, if user evaluations fail to register even major performance failures, the utility of user satisfaction surveys as tools for holding public service providers accountable becomes questionable. This is particularly so if the resources needed to monitor user satisfaction are taken from more traditional hierarchical oversight and accountability procedures (Damgaard and Nielsen 2020), including more informal channels for feedback that managers may also respond to (Kroll 2013; Olsen 2017). This concern about subjective performance indicators is anything but new (Stipak 1979), and studies clearly show how satisfaction and evaluation measures are overly sensitive, picking up factors that are seemingly irrelevant (Olsen 2015; Thau et al. 2021; Van de Walle and Van Ryzin 2011). We add to existing evidence by showing that satisfaction as an indicator is not always too sensitive but can also be too insensitive.

Third, the fact that service users do not voice their dissatisfaction even when they have plausibly been directly hurt by bureaucratic malpractice seems to question the understanding of satisfaction surveys as an instrument that empowers users by feeding their experiences back into the service delivery process (Hjortskov 2017; Kelly 2005). Our findings could be taken to suggest that service users are simply unable to play the role of ensuring accountability that is often envisioned for them. Yet, such a conclusion would be premature. For one thing, the results are specific to our case, and additional evidence is needed to investigate whether they can be replicated. Furthermore, in our view, the main question going forward is not whether but *under which conditions* user satisfaction is responsive to performance failures.

10 | Limitations

Our findings should be interpreted in light of several limitations. First, our research design does not allow us to directly assess how respondents processed the information about the NBSS performance failure. Despite the widespread media coverage of the scandal, we cannot be certain that NBSS users had the failure in mind when completing the satisfaction survey. Consequently, we cannot disentangle whether the null finding is due to respondents actively dismissing the failure, incrementally updating their perceptions, attributing blame to a single employee, or simply being inattentive to the event. This will be a limitation inherent to most studies examining real-world performance failures—unless respondents are explicitly primed with unambiguous information immediately before measurement. However, such priming introduces other trade-offs, distancing the research from the conditions under which satisfaction judgments are typically formed.

Second, our survey relies on individual respondents answering on behalf of their organizations. This may introduce some heterogeneity, as contact persons and organizations may have followed different internal processes for completing the survey—whether responding individually, collaboratively, or delegating to another colleague. While we cannot observe how these processes varied across organizations, we have no reason to believe it introduces systematic bias into our identification strategy, as the same conditions apply both before and after the performance failure. Still, this heterogeneity could introduce noise into our data, reducing the precision of our estimates.

Third, it is possible that the performance failure might have had effects on user assessments other than satisfaction with NBSS's grant management. Satisfaction has a central place in public administration theory and practice, especially in relation to performance management and citizen involvement (Moynihan 2008; Van Ryzin 2006), but service users possibly make differentiated evaluations. While the NBSS performance failure arguably signaled low performance, weak procedures, and unethical conduct in the context of grant management, some users may have viewed the satisfaction survey primarily as a venue for expressing individualistic experiences regarding the ease and professionalism of the grant management system. NBSS users could have changed moral or procedural perceptions of the organization's reputation (Maor et al. 2025; Overman et al. 2020) following the scandal but not viewed these attitudes as relevant to the survey's satisfaction questions. Although such a differentiation requires a level of sophistication that may be unrealistic for typical user satisfaction surveys, further work should examine this by distinguishing various dimensions of service evaluations.

Qualitative interviews with users may also be well-suited to exploring such questions.

Fourth, all respondents in our sample had sufficient experience and competence to successfully apply for NBSS grants and were ultimately awarded funding. It is therefore possible that the effects of the performance failure would have differed among unsuccessful applicants, who are presumably less competent, on average. While we find that less competent users indeed report lower satisfaction with the NBSS overall, our analyses do not indicate that (self-reported) user competence moderated the impact of the performance failure: for highly and less competent organizations alike, effects are insignificant (SI, Section B).

Fifth, another characteristic of the NBSS case is that we are dealing with a government agency that is the sole provider of a specific service (i.e., grant management and payouts in the social policy area). The NBSS users thus have no alternative providers to choose from, while at the same time being completely dependent on NBSS services (i.e., funding decisions). This lack of an exit option is generally thought to make the voice option more salient to service users (Hirschman 1970). Yet, when the organizational survival of NBSS users hinges on year-to-year funding, one could expect NBSS users to be less eager to signal discontent. Respondents were, however, not uncritical of the NBSS, with average satisfaction around 6.5 (1–10 scale), and anonymity made expressing dissatisfaction on the survey low-risk. Still, studies on scandal-level performance failures in non-monopoly situations would help to uncover the trade-offs that individual users face in voicing discontent.

Finally, we encourage studies that consider the effect of performance failures outside a high-trust, low-corruption context. While we argued that Denmark provided a most-likely case in which to observe a negative effect of performance failure on user satisfaction, an alternative view is that the high-trust, low-corruption context is actually a least-likely case. For example, with a generally uncorrupt bureaucracy, the scale of the NBSS failure could lead service users to treat it as an outlier, decoupling it from the everyday workings of the NBSS. Also, high institutional trust could lead to a blind faith that the political system self-corrects performance failures, making it unnecessary for service users to express their discontent.

11 | Conclusion

This article challenges assumptions in the satisfaction literature, and in public administration practice in many countries, regarding a service feedback system where users respond to low performance or poor service quality by expressing dissatisfaction when observed performance falls below expectations. Using a natural experiment, we find no significant negative effect of one of the most visible and dramatic performance failures in Denmark in recent memory. This null finding is robust across various model specifications, including several subgroup analyses, and none of our tests of key underlying assumptions—regarding balance on co-variates, pre-existing trends, and response rates before and after the NBSS scandal—indicate concerns. However, null findings are inherently challenging as they can be the result of contextual peculiarities. Could our findings be due to the specific nature of our respondents, that they were inattentive to the performance failure, or that they attributed the failure solely to an individual rather than the organization?

While we addressed these concerns in detail above, we return to them here to make a more fundamental point: If a theory can only explain outcomes under idealized conditions (e.g., in an experiment), its practical utility for understanding real-world behavior is limited. In our case, it is certainly possible that some respondents deemed the performance failure not salient to their evaluation for one reason or another. Yet, this is not a unique limitation of our study. Rather, it reflects the very phenomenon we wish to understand: how service recipients use (or fail to use) performance information in real-world evaluations. Indeed, in many settings, citizens and users are likely to be inattentive, hold rigid attitudes, or be inclined to assign blame to specific individuals, circumstances, or external forces. This lack of responsiveness to performance signals is not necessarily an indication that user evaluations are nonsensical but can be explained by motivated reasoning (a type of cognitive bias) or strongly held priors (under a Bayesian model of satisfaction) – factors for which the EDM does not adequately account.

For practitioners, our findings serve as a reminder to take quantitative satisfaction data (like any other single performance metric) with a grain of salt. Satisfaction surveys provide an important opportunity for users to express opinions about an organization, but the survey format is also constraining in some ways, and users are subject to limitations in their cognitive capacities and access to information, just like all other individuals. As noted previously, some respondents in our study did explicitly react to the unfolding scandal in their responses to open-ended survey questions, despite the lack of any detectable drop in quantitative satisfaction scores. This points to the importance of open-ended questions, and other inputs like qualitative interviews, as valuable sources of information from service users. Although such data pose challenges when it comes to easily summarizing aggregate responses, a holistic approach to service feedback is, in our view, worth the extra effort. Despite our findings, we still believe quantitative user satisfaction data can be a useful tool. Like any metric, however, it has limitations and should be interpreted in context alongside other signals of performance available to managers and policy makers.

Acknowledgments

This work has been presented at the Danish Political Science Association's 2021 conference and the European Group for Public Administration's 2022 conference. We appreciate the helpful comments received on these occasions. We also wish to thank Ulrik Hvidman, Rasmus Tue Pedersen, and Gregg Van Ryzin for valuable feedback.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Due to the confidentiality of participants, the data used cannot be made publicly available. Code and script will therefore not be deposited publicly but can be shared upon request. Our pre-analysis plan is available at https://doi.org/10.17605/OSF.IO/ZRXUK. Due to language editing,

the wording of the hypothesis in the article and preregistration differ slightly, but the meaning is identical.

Endnotes

[1] Our pre-analysis plan identifies additional sources of potential effect heterogeneity, as described in our results.

[2] See https://doi.org/10.17605/OSF.IO/ZRXUK. We were 'partially' rather than 'fully' blinded (Nosek et al. 2018, 2603). The data already existed at the time of registration, and the distribution on user satisfaction in 2018 and 2020 had been observed. However, the effect of the embezzlement scandal had not been estimated, and we thus selected hypotheses and an analysis plan while blind to the results.

[3] Box plots of how the satisfaction distributions changed during the collection period in 2018 (and 2020) are found in the SI, Section A. Here, we also plot and elaborate on response patterns in the collection periods.

[4] We experimented with various windows around the scandal and found that these two periods struck a balance between sufficient observations and a narrow window. The restricted before-and-after periods include, respectively, 27% and 25% of the full sample. The other periods tried out did not lead to other conclusions.

[5] We note that several other variables in the models are significant and in the expected directions (e.g., competencies).

[6] For example, highly satisfied respondents might be more eager to participate in the survey immediately than less satisfied respondents, who might need a reminder or two before participating.

[7] The ratings relate to the general call for applications (announcement, guidelines, purpose statement) and the decision letter sent to applicants (readability, justification given, payment instructions).

References

Andersen, S. C., and M. Hjortskov. 2016. "Cognitive Biases in Performance Evaluations." *Journal of Public Administration Research and Theory* 26, no. 4: 647–662.

Andrews, R., G. A. Boyne, and G. Enticott. 2007. "Performance Failure in the Public Sector: Misfortune or Mismanagement?" *Public Management Review* 8, no. 2: 273–296.

Baekgaard, M., and S. Serritzlew. 2016. "Interpreting Performance Information: Motivated Reasoning or Unbiased Comprehension." *Public Administration Review* 76, no. 1: 73–82.

Brehm, J., and S. Gates. 1997. *Working, Shirking, and Sabotage. Bureaucratic Response to a Democratic Public*. University of Michigan Press.

Brown, K., and P. B. Coulter. 1983. "Subjective and Objective Measures of Police Service Delivery." *Public Administration Review* 43, no. 1: 50–58.

Damgaard, P. R., and P. A. Nielsen. 2020. "Does Performance Disclosure Affect User Satisfaction, Voice, and Exit? Experimental Evidence From Service Users." *Journal of Behavioral Public Administration* 3, no. 2. https://doi.org/10.30636/jbpa.32.113.

Ditzel, E. E. 2021. *Britta. Forræderiet Mod Velfærdsstaten*. People's.

Djerf-Pierre, M., M. Ekström, and B. Johansson. 2013. "Policy Failure or Moral Scandal? Political Accountability, Journalism and New Public Management." *Media, Culture and Society* 35, no. 8: 960–976.

DR. 2018. "Millionsvindel i Socialstyrelsen: Det Ved vi Om Sagen [What We Know About the Embezzlement in NBSS]." https://www.dr.dk/nyheder/indland/millionsvindel-i-socialstyrelsen-det-ved-vi-om-sagen.

Favero, N., and M. Kim. 2021. "Everything Is Relative: How Citizens Form and Use Expectations in Evaluating Services." *Journal of Public Administration Research and Theory* 31, no. 3: 561–577.

Favero, N., R. M. Walker, and J. Zhang. 2025. "A Dynamic Study of Citizen Satisfaction: Replicating and Extending Van Ryzin's "Testing the Expectancy Disconfirmation Model of Citizen Satisfaction With Local Government"." *Public Management Review* 27, no. 6: 1588–1606.

Fornell, C., M. D. Johnson, E. W. Anderson, J. Cha, and B. E. Bryant. 1996. "The American Customer Satisfaction Index: Nature, Purpose, and Findings." *Journal of Marketing* 60, no. 4: 7–18.

Gottschalk, P. 2021. "11. Social Security by PwC." In *Private Policing of Economic Crime: Case Studies of Internal Investigations*. Routledge.

Grøn, C. H., and M. B. Kristiansen. 2022. "What Gets Measured Gets Managed? The Use of Performance Information Across Organizational Echelons in the Public Sector." *Public Performance & Management Review* 45, no. 2: 448–472.

Hirschman, A. 1970. *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Harvard University Press.

Hjortskov, M. 2017. "Priming and Context Effects in Citizen Satisfaction Surveys." *Public Administration* 95, no. 4: 912–926.

Hjortskov, M. 2019. "Citizen Expectations and Satisfaction Over Time: Findings From a Large Sample Panel Survey of Public School Parents in Denmark." *American Review of Public Administration* 49, no. 3: 353–371.

James, O., and A. Moseley. 2014. "Does Performance Information About Public Services Affect Citizens' Perceptions, Satisfaction, and Voice Behaviour? Field Experiments With Absolute and Relative Performance Information." *Public Administration* 92, no. 2: 493–511.

Jensen, J. K., M. Thau, and M. F. Mikkelsen. 2021. "Tilfredshedsundersøgelse Blandt Socialstyrelsens Tilskudsmodtagere." www.vive.dk.

Jilke, S., and M. Baekgaard. 2020. "The Political Psychology of Citizen Satisfaction: Does Functional Responsibility Matter?" *Journal of Public Administration Research and Theory* 30, no. 1: 130–143.

Jilke, S., and L. Tummers. 2018. "Which Clients Are Deserving of Help? A Theoretical Model and Experimental Test." *Journal of Public Administration Research and Theory* 28, no. 2: 226–238.

Kelly, J. M. 2003. "Citizen Satisfaction and Administrative Performance Measures: Is There Really a Link?" *Urban Affairs Review* 38, no. 6: 855–866.

Kelly, J. M. 2005. "The Dilemma of the Unsatisfied Customer in a Market Model of Public Administration." *Public Administration Review* 65, no. 1: 76–84.

Kroll, A. 2013. "The Other Type of Performance Information: Nonroutine Feedback, Its Relevance and Use." *Public Administration Review* 73, no. 2: 265–276.

Kunda, Z. 1990. "The Case for Motivated Reasoning." *Psychological Bulletin* 108, no. 3: 480–498.

Maor, M., D. Rimkutė, and T. Capelos. 2025. "Emotions and Reputation Learning by Audience Networks: A Research Agenda in Bureaucratic Politics." *Public Administration Review*. https://doi.org/10.1111/puar.70004.

Marvel, J. D. 2016. "Unconscious Bias in Citizens Evaluations of Public Sector Performance." *Journal of Public Administration Research and Theory* 26, no. 1: 143–158.

Moynihan, D. P. 2008. *The Dynamics of Performance Management: Constructing Information and Reform*, edited by D. P. Moynihan. Georgetown University Press.

Moynihan, D. P., and D. P. Hawes. 2012. "Responsiveness to Reform Values: The Influence of the Environment on Performance Information Use." *Public Administration Review* 72, no. 1: S95–S105.

Mulgan, R. 2000. "'Accountability': An Ever-Expanding Concept?" *Public Administration* 78, no. 3: 555–573.

Muñoz, J., A. Falcó-Gimeno, and E. Hernández. 2020. "Unexpected Event During Survey Design: Promise and Pitfalls for Causal Inference." *Political Analysis* 28, no. 2: 186–206.

Nosek, B. A., C. R. Ebersole, A. C. DeHaven, and D. T. Mellor. 2018. "The Preregistration Revolution." *Proceedings of the National Academy of Sciences* 115, no. 11: 2600–2606.

Oliver, R. L. 1980. "A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions." *Journal of Marketing Research* 17, no. 4: 460–469.

Olsen, A. L. 2015. "Citizen (Dis)satisfaction: An Experimental Equivalence Framing Study." *Public Administration Review* 75, no. 3: 469–478.

Olsen, A. L. 2017. "Human Interest or Hard Numbers? Experiments on Citizens' Selection, Exposure, and Recall of Performance Information." *Public Administration Review* 77, no. 3: 408–420.

Overman, S., M. Busuioc, and M. Wood. 2020. "A Multidimensional Reputation Barometer for Public Agencies: A Validated Instrument." *Public Administration Review* 80, no. 3: 415–425.

Schachter, H. L. 2010. "Objective and Subjective Performance Measures: A Note on Terminology." *Administration and Society* 42, no. 5: 550–567.

Song, M., S.-H. An, and S. G. S. Yang. 2025. "Socioeconomic Disparities, Service Equity, and Citizen Satisfaction: Cross-National Evidence." *Public Administration Review* 85, no. 4: 973–988.

Song, M., and K. J. Meier. 2018. "Citizen Satisfaction and the Kaleidoscope of Government Performance: How Multiple Stakeholders See Government Performance." *Journal of Public Administration Research and Theory* 28, no. 4: 489–505.

Stipak, B. 1979. "Citizen Satisfaction With Urban Services Potential Misuse as a Performance Indicator." *Public Administration Review* 39, no. 1: 46–52.

Thau, M., M. F. Mikkelsen, M. Hjortskov, and M. J. Pedersen. 2021. "Question Order Bias Revisited: A Split-Ballot Experiment on Satisfaction With Public Services Among Experienced and Professional Users." *Public Administration* 99, no. 1: 189–204.

Transparency International. 2020. "Corruption Perceptions Index 2020. Tranparency International. The Global Coalition Againt Corruption. 2020." https://www.transparency.org/en/cpi/2020/table/dnk#.

Van de Walle, S., and G. G. Van Ryzin. 2011. "The Order of Questions in a Survey on Citizen Satisfaction With Public Services: Lessons From a Split-Ballot Experiment." *Public Administration* 89, no. 4: 1436–1450.

Van Ryzin, G. G. 2006. "Testing the Expectancy Disconfirmation Model of Citizen Satisfaction With Local Government." *Journal of Public Administration Research and Theory* 16, no. 4: 599–611.

Van Ryzin, G. G. 2013. "An Experimental Test of the Expectancy-Disconfirmation Theory of Citizen Satisfaction." *Journal of Policy Analysis and Management* 32, no. 3: 597–614.

Zhang, J., W. Chen, N. Petrovsky, and R. M. Walker. 2022. "The Expectancy-Disconfirmation Model and Citizen Satisfaction With Public Services: A Meta-Analysis and an Agenda for Best Practice." *Public Administration Review* 82, no. 1: 147–159.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1:** puar70059-sup-0001-Supinfo.pdf.

Biographies

Mads Thau is Senior Researcher at the Institute for Social Research in Oslo and has a PhD in political science from Aarhus University. His research interests cover democratic politics, civil society, local government, and public administration. For example, Thau has published on public service delivery, forms of government, politicians' working conditions, teamwork among professionals, client-bureaucrat encounters, and citizen and user satisfaction.

Maria Falk Mikkelsen is Senior Researcher at VIVE – The Danish Center for Social Science Research and Associate Professor at the Department of Political Science, University of Southern Denmark. She holds a PhD in Political Science from Aarhus University. Her research focuses on policy implementation, performance measurement, and decision-making in the education and welfare sectors. Her current research projects focus on social equity and on how citizens and service users evaluate public services and policies.

Nathan Favero is Provost Associate Professor in the Department of Public Administration & Policy within the School of Public Affairs at American University. He received a PhD in political science from Texas A&M University in 2016. His research interests include citizen satisfaction, performance, management, social equity, quantitative methodology, education policy, and systems-level models of public administration/policy.