Journal of Integrated Knowledge
Vol. 2 No. 11 (2026)

Technical Quality and Engagement: Reading Observational and Quasi-Experimental Evidence Together

Brandon Smith[1] · Zachary Stevens[2] · Benjamin Lewis[3]

[1] Nauru · [2] Egypt · [3] China

Published: November 05, 2026 | Article 943

Abstract

Different observational designs can address different parts of the quality-engagement relationship. Their findings should be compared through their targets and assumptions, rather than treated as interchangeable confirmations.

Technical Quality and Engagement: Reading Observational and Quasi-Experimental Evidence Together

Abstract

Different observational designs can address different parts of the quality-engagement relationship. Their findings should be compared through their targets and assumptions, rather than treated as interchangeable confirmations.

Introduction

The association question in Yuan et al. (2026) can be placed alongside broader evidence about playback and engagement without assuming that every design estimates the same quantity. A descriptive contrast characterizes the observed record. A design intended to support a stronger causal interpretation introduces additional assumptions and a different evidential target. A review should identify that target before comparing the signs or magnitudes of findings. Disagreement may then reflect a change in the question rather than a failure of replication. The value of a mixed evidence base lies in the distinctions it can clarify, including the assumptions that remain unsupported across all available designs.

Association and intervention require different evidence

The quasi-experimental approach in Krishnan and Sitaraman (2013) focuses on identifying patterns associated with specific behaviors, such as rapid exits from a system. In contrast, observational studies like those in Shmueli (2010) aim to explain underlying mechanisms and the conditions under which certain outcomes occur. While both can reveal associations between technical factors and engagement, the nature of their findings differs significantly. A model that identifies patterns associated with rapid exits, for instance, does not necessarily provide insight into what would happen if a technical condition were altered. This highlights the importance of distinguishing between association studies, which focus on identifying correlations, and intervention designs, which aim to understand causal relationships.

Future research should target these missing links rather than simply describe every predictive improvement as an intervention benefit. By focusing on the specific evidence each type of study provides-technical context for observed behaviors versus design-based arguments about changes-the coherence of the streaming topic is maintained without a universal interpretation of engagement. This approach ensures that the distinctions between observational and quasi-experimental designs are respected, allowing for a more nuanced understanding of the quality-engagement relationship.

Quality and engagement evidence

Prediction attribution, as described by Lundberg & Lee (2017), focuses on explaining how models interpret their inputs under specific conditions. This explanation is distinct from the actual process that generates an event, which is a separate evidential object. In contrast, precision-recall analysis, outlined by Saito & Rehmsmeier (2015), evaluates model performance within a defined context and event definition. A high discrimination summary does not automatically imply a favorable operating threshold or the outcomes of false alerts. These are critical decisions that should be made explicitly.

Neural-network calibration, as discussed in Guo et al. (2017), distinguishes between confidence measures and the frequency of correct predictions for a specific outcome. Calibration is conditional on both an event and its population, meaning that a probability assigned to a reported label does not automatically

become a probability for an unrelated underlying construct. These distinctions highlight the need to carefully examine each type of evidence in relation to its specific context and assumptions.

Session dependence changes the intended test

Session dependence can alter the intended test when examining observational designs, as evidenced by Kapoor and Narayanan (2023) and Koh et al. (2020). These studies highlight that what is held out-whether it's a session or an exposure-is crucial for understanding the quality of engagement. A playback dataset might contain numerous exposures from the same session, making a held-out exposure more familiar than one from a new session. This familiarity can skew results, as the environment perspective in Koh et al. (2020) and the leakage perspective in Kapoor and Narayanan (2023) emphasize. Therefore, it is essential to clarify what was actually held out and how this affects the validity of the findings. Additionally, the selection of adjudicated examples can also influence the interpretation, as heavily reviewed cases might not represent typical traffic. Thus, evaluating these examples carefully ensures that the study's conclusions are more reflective of the known population rather than assuming broad generality from large observational coverage.

Weak labels and predictive evaluation

The leakage analysis by Kapoor and Narayanan (2023) highlights that even with large datasets and sophisticated models, boundary errors between model development and evaluation can still occur. These errors result from a lack of clarity about what information was available at each decision point and whether the assessment remained independent of selection criteria. This is particularly relevant when comparing observational designs, as they may target different aspects of the quality-engagement relationship.

WILDS (Koh et al., 2020) provides a benchmark perspective on changes between training and evaluation environments. The key issue here is the need to clearly define the destination population and recognize that a new input distribution, altered labeling relationships, or revised tasks can each present unique challenges. These differences should not be conflated into a single robustness claim, as they represent distinct issues requiring separate scrutiny.

The distinction between explaining and predicting, as articulated by Shmueli (2010), underscores why a model that fits data well does not necessarily explain the underlying mechanism or causal relationship. A review can value predictive success while maintaining uncertainty about the causality or substantive explanation of the observed relationship. This is crucial when comparing observational designs, which may aim to predict engagement based on various quality factors without necessarily explaining the mechanisms at play.

Weak labels have an availability time

The weak-supervision perspective in Ratner et al. (2017) acknowledges the utility of imperfect labels but highlights their limitations, particularly concerning the availability time and predictive value. Ratner et al. distinguish between a technical event, a later adjudication, and a behavioral summary, noting that these different types of information have varying availability times. This means that a retrospective record might contain details that a real-time prediction cannot access, such as outcomes or decisions made after the initial observation. Kapoor and Narayanan (2023) further reinforce this by emphasizing the importance of distinguishing between signals used to define a label and those permitted as predictors at the decision point. They argue that combining labels does not make them independent. If they arise from one observation or reporting pathway, their interdependence remains significant. Therefore, a future evaluation should clearly state which event the score predicts and what evidence was genuinely available before that event was known.

Confidence and contextual interpretation

In the context of the quasi-experimental video-quality study by Krishnan and Sitaraman (2013), confidence in findings hinges on a clear understanding of the design's objectives and assumptions. Unlike predicting labels from historical logs, which relies on retrospective analysis, the study explicitly frames a causal-inference question about viewing behavior. This distinction underscores that different observational designs are not interchangeable confirmations but rather address distinct parts of the quality-engagement relationship.

Similarly, Snorkel’s approach to creating training data from weak supervision (Ratner et al., 2017) highlights the importance of contextual interpretation in observational studies. Multiple labeling sources can be valuable, but their reliance on imperfect and dependent data necessitates a synthesis that examines the origins of each label. If apparent agreement among labels is due to independent support rather than a common reporting pathway, this should be acknowledged. A synthesis must explain how observed values relate to the complete population, ensuring that confidence in findings is not misplaced.

The missing-data framework (RUBIN, 1976) emphasizes the central role of assumptions about the observation process in inference. Just because a model accepts incomplete records does not automatically render these observations harmless. A synthesis should elucidate how observed values reflect the complete population described in the conclusion, thereby maintaining the integrity of the findings.

The evaluation unit follows the intended destination

When evaluating exposures that share a session, device, account context, and recording process, the number of rows alone does not sufficiently identify independent information relevant to a claim about engagement or technical quality. This is because the environment-shift perspective proposed by Koh et al. (2020) is particularly valuable when assessing changes to new technical or behavioral contexts rather than merely comparing known exposures. For instance, if an exposure involves a novel interaction with a device that alters user behavior in ways not seen before, this shift necessitates a more nuanced understanding of the data. In such cases, the focus should be on how the new context impacts engagement and quality, rather than simply confirming existing findings from similar exposures. This distinction is crucial for developing a robust evaluation unit that accurately reflects the intended destination and provides meaningful insights into how user behavior changes under different conditions.

A threshold is chosen under an error tradeoff

In the precision-recall perspective articulated by Saito and Rehmsmeier (2015), the tradeoff between precision and recall is crucial for evaluating discrimination, especially under class imbalance conditions. Similarly, a threshold choice reflects the balance between false positives and false negatives, impacting the practical significance of an alert system. When comparing observational designs, it's essential to recognize that different thresholds can have divergent implications based on the specific context and assumptions underlying each design. For instance, in community content reporting, a higher threshold might be more appropriate for detecting significant events, whereas in media engagement studies, a lower threshold could better capture minor but frequent interactions. Thus, rather than treating these designs as interchangeable confirmations, it's necessary to assess how changes in thresholds affect the operational outcomes and their practical relevance.

To mitigate the risk of overlooking these nuances, a future evaluation should retain both the threshold-selection data and the independent assessment population. By distinguishing between changes in event prevalence and shifts in the relationship between observed signals and labels, researchers can better contextualize the findings and maintain the necessary uncertainty. This approach allows for a more nuanced understanding of how different designs contribute to the quality-engagement relationship, preserving the complexity that remains after a model issues a score.

Discrimination does not select the service response

Saito and Rehmsmeier (2015) and Guo et al. (2017) propose contrasting approaches to understanding technical quality and engagement, each addressing distinct aspects of the relationship. Saito and Rehmsmeier focus on distinguishing cases, questioning the appropriateness of certain actions based on observed data. Meanwhile, Guo et al. (2017) concentrate on the meaning of issued probabilities in relation to defined events. Neither study independently selects a threshold that aligns with service responses. This choice is influenced by the outcomes and associated burdens of errors.

In future assessments, it is crucial to retain these threshold-selection procedures and the populations for which they were originally designed. As the prevalence or technical context changes, the same operating point may carry different meanings. Moreover, it is essential to differentiate between alerts that prompt review and automatic interventions, as the evidence and costs required to justify these responses can vary significantly.

Uncertain feedback should remain visible to a response

In Ratner et al.'s (2017) work, the weak-supervision perspective encourages viewing labels as nuanced rather than absolute. This approach acknowledges that some cases might not be definitively classified, necessitating a cautious handling of feedback. In future evaluations, it's crucial to document which cases receive further scrutiny and what new insights emerge from these reviews. It's important to differentiate between justified deferrals-where additional information is sought-and avoidances of complex or ambiguous scenarios. While model explanations by Lundberg and Lee (2017) can assist in inspection, they don't necessarily provide definitive answers about intent or intervention benefits. By integrating uncertainty management into playback analysis, we can prevent a calibrated score for one label from becoming an unqualified claim about the underlying reason or action. This practice ensures that the ambiguous feedback remains visible, allowing for a more nuanced response to observational and quasi-experimental evidence.

What the observation represents

In a hypothetical chart representing media engagement, one entry might show the number of articles published on local community content by a single news outlet over a week. This unit captures the aggregate output of a single entity, providing insight into the quantity and quality of reporting on specific topics.

At a second level of aggregation, this same data could represent the total number of articles published on community content by all outlets within a region in a week. Here, the unit shifts to aggregate regional media coverage, giving context to broader patterns and trends across multiple sources. The interpretation at this level would focus more on the overall impact and reach of community-focused reporting, rather than the specific output of individual outlets.

Sorting records into comparable groups

Sorting records into comparable groups can influence the observed association between quality and engagement. For instance, consider a study examining the impact of media content quality on audience engagement. If researchers group participants based on their initial levels of interest or engagement, those who are already more engaged might be disproportionately represented in higher-quality content groups. This could exaggerate the positive effect of high-quality media on engagement, as it may be harder for lower-engaged participants to shift their attention to engaging with such material. Alternatively, if researchers group participants based on their frequency of media consumption, those who consume media less often might be overrepresented in lower-quality content groups. This could understate the impact of higher quality media, as these participants might not have sufficient exposure to engage deeply. Both scenarios suggest that grouping participants into similar groups can alter the observed association between the variables of interest, highlighting the importance of careful selection criteria in observational studies.

Conclusion

One consequence of this approach is that it highlights the importance of documenting and distinguishing between cases that require further scrutiny and those that are definitively classified. This prevents the ambiguous feedback from being treated as definitive evidence, ensuring a more nuanced interpretation of observational and quasi-experimental evidence. An open question is how to effectively quantify and compare the impact of different observational designs, especially when they target different aspects of the quality-engagement relationship. Another challenge is how to integrate uncertainty management into the interpretation of feedback, particularly when labels are nuanced rather than absolute.

References

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., et al. (2020). WILDS: A Benchmark of in-the-Wild Distribution Shifts. *arXiv preprint arXiv:2012.07421*. https://arxiv.org/abs/2012.07421

Shmueli, G. (2010). To Explain or to Predict?. *Statistical Science, 25(3), 289-310*. https://doi.org/10.1214/10-STS330

Kapoor, S., Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns, 4(9), 100804*. https://doi.org/10.1016/j.patter.2023.100804

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., Ré, C. (2017). Snorkel: Rapid Training Data Creation with Weak Supervision. *arXiv preprint arXiv:1711.10160*. https://arxiv.org/abs/1711.10160

Lundberg, S., Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. *arXiv preprint arXiv:1705.07874*. https://arxiv.org/abs/1705.07874

Yuan, Y., Xu, T., Huang, J., & Yin, J. (2026). Playback Quality Anomalies Associated with Rapid Swiping in Short Videos: Model-attributed Association Estimation and Risk Prediction. *SSRN working paper 7393398*. doi:10.2139/ssrn.7393398.

Krishnan, S. S., Sitaraman, R. K. (2013). Video Stream Quality Impacts Viewer Behavior: Inferring Causality Using Quasi-Experimental Designs. *IEEE/ACM Transactions on Networking, 21(6), 2001-2014*. https://doi.org/10.1109/tnet.2013.2281542

RUBIN, D. B. (1976). Inference and missing data. *Biometrika, 63(3), 581-592*. https://doi.org/10.1093/biomet/63.3.581

Saito, T., Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE, 10(3), e0118432*. https://doi.org/10.1371/journal.pone.0118432