Annals of Science and Inquiry
Vol. 2 No. 11 (2026)

Optimizing a Reaction Mean while Accounting for Batch Variation

Austin Stone[1] · David Hayes[2] · Gavin Bell[3]

[1] Haiti · [2] Myanmar · [3] Hungary

Published: November 02, 2026 | Article 991

Abstract

The condition with the highest estimated mean may not be the most dependable choice across batches. Optimization should distinguish the average response from variation relevant to a new batch.

Optimizing a Reaction Mean while Accounting for Batch Variation

Abstract

The condition with the highest estimated mean may not be the most dependable choice across batches. Optimization should distinguish the average response from variation relevant to a new batch.

Introduction

The mixed-model optimization topic of Zhao et al. (2023) can target a population mean while a practical decision concerns a future batch. Those targets differ because batch variability need not disappear at the estimated mean optimum. A review should identify which response is being optimized and what uncertainty accompanies the recommendation. A narrow interval for an average is not automatically a narrow range for a new realization. The model's grouping structure helps express that distinction only when it follows the actual design. A recommendation becomes more informative when its expected benefit and its batch-level uncertainty are both attached to the condition being recommended.

Mixed-model structure expresses a scientific assumption

In Bates et al.'s (2015) computational framework and Gelman and Hill's (2006) multilevel perspective, a mixed-model structure is proposed to express the scientific assumption that structured variation exists within the data. This structure acknowledges that observations are not independent but are influenced by systematic differences, such as batch effects or experimental conditions. The choice of grouping factors, covariance structures, and pooling strategies in these models should reflect the context in which the data were collected. A model's effectiveness is limited if it does not account for the specific production process or population of interest. Thus, a mixed-model structure must be grounded in the data-generating mechanism to accurately represent the variability among groups.

Moreover, a well-fitted model does not guarantee that the underlying assumptions are correct or that the conclusions drawn from it are reliable. A sensitivity analysis can help identify which model structures are robust and which might lead to different interpretations of the results. This approach makes model selection an explicit part of the inference process, rather than an implicit decision made by default in software settings. By examining how different modeling choices affect the substantive findings, researchers can better justify their conclusions based on the data and the research question at hand.

Design and response-surface evidence

The distinction between explaining and predicting underscores why a fitted model might not reveal the underlying mechanism (Shmueli, 2010). A review should prioritize its inferential goal over model interpretation. While predictive success is valuable, uncertainty remains about the causal or substantive explanation of observed relationships.

The lme4 paper offers a computational framework for fitting linear mixed-effects models (Bates et al., 2015). However, successful estimation alone does not confirm the chosen model structure answers the scientific question. Instead, the structure should align with the allocation, dependence, and intended generalization rather than convenience of available software.

Reproducibility has several possible objects

Reproducing a measurement and achieving consistent patterns are distinct challenges. While agreement within a single batch enhances confidence in its consistency, it does not guarantee reproducibility across different batches. To develop reproducibility, we must distinguish between the average response and the variation relevant to new batches. A detailed record of what was held constant and what changed should be provided, along with preserving discordant outcomes with sufficient context. Such outcomes can help identify the boundary of an explanation rather than simply count as failed replications. This approach gives future work a clearer purpose by focusing on the uncertainty each comparison reduces.

Adaptive experimentation creates a sequence of claims

Adaptive experimentation suggests selecting conditions based on initial outcomes rather than predetermined comparisons, aligning with the notion that the highest mean response might not be the most reliable across different batches. This approach emphasizes distinguishing between average performance and variation relevant to new conditions, as indicated by the mixed-model structure and design-response-surface evidence. For instance, if a reaction's mean output appears optimal in early trials, it does not guarantee this will hold true in subsequent batches due to inherent variability. Therefore, an adaptive strategy should focus on identifying which comparisons are valid for the current batch and avoid presenting all choices as pre-determined. A future report could detail this sequence by acknowledging unfavorable conditions that led to the chosen recommendation, thereby making the optimization process reproducible as a decision pathway. The scientific conclusion might then address the selected condition, the adaptive procedure itself, or both, supported by evidence fitting each perspective.

Structured variation and estimation

Structured variation and estimation emphasize how data from multiple batches should be analyzed to avoid misleading conclusions. Pooling techniques, such as those described by Gelman and Hill (2006), help stabilize estimates by considering the overall pattern of the data rather than focusing solely on local variations within each batch. This approach acknowledges that the condition with the highest estimated mean might not be the most dependable choice across different batches. Instead, it recognizes that the average response must be distinguished from variation specific to a new batch.

The response-surface methodology, as outlined by Box and Wilson (1951), provides a framework for experimental optimization. Its scope is limited by the design region and the observations used to estimate the relationship between conditions and outcomes. A numerical optimum derived through optimization does not guarantee that extrapolation or future use will be supported, as it may not account for the variability present in new batches. This highlights the importance of considering structured variation when optimizing a reaction mean, ensuring that estimates are robust and reliable across different experimental settings.

A new outcome has more uncertainty than a fitted mean

The random-effects framework, as described by Laird and Ware (1982), and the small-sample perspective outlined by Kenward and Roger (1997), highlight the distinction between uncertainty about an average response and variability specific to a future outcome. When considering a new batch, the decision to adopt a recommended condition might be misguided if it relies on variability observed within existing batches. A narrow interval for the fitted mean does not provide enough information to assess whether this recommendation is appropriate for a new unit. The model can estimate the average relationship and characterize this part of the decision well, but it leaves broader predictions limited by the design's structure.

Thus, it becomes crucial to explicitly distinguish between the uncertainty about an average response and the variability relevant to a future batch. This distinction allows the apparent precision of an optimal choice to be evaluated without dismissing the value of the fitted model. By acknowledging the assumptions that remain untested, we can ensure that decisions are based on more than just the estimated mean, thereby enhancing the reliability of optimization strategies across different batches.

Inference and the observation process

The work on small-sample inference under fixed effects and restricted maximum likelihood (Kenward & Roger, 1997) highlights the importance of considering uncertainty beyond asymptotic approximations. This is particularly relevant when comparing different batches or conditions, as a single batch's highest mean response may not be the most dependable choice across all samples. The applicable uncertainty in this context depends on the design of the experiment, the fitted model, and the population it represents.

The variance-component approximation (Satterthwaite, 1946) underscores that the numerical precision of an estimated uncertainty quantity does not equate to its sampling behavior's certainty. An inferential statement must therefore align with the units of independent information supporting it. This means distinguishing between average response and variation specific to a new batch.

False-discovery-rate control (Benjamini & Hochberg, 1995) is another concern when making multiple statistical comparisons. It does not transform a selected favorable condition into independent confirmation. Instead, it emphasizes the need for rigorous methods to avoid false positives in an adaptive experimentation sequence designed to optimize and account for batch variation.

A fitted optimum belongs to a design region

The response-surface perspective in Box and Wilson (1951) and the quantitative design perspective in Box and Behnken (1960) suggest that an estimated optimum corresponds to the conditions actually studied, within the scope of the provided evidence. Numerical optimization, however, may yield a point outside this region. The precision of the numerical result does not necessarily reflect the precision of the scientific conclusion, which is often more about the average response and reproducibility across batches rather than the specific conditions used in the study.

The argument here is that a fitted optimum belongs to a design region, meaning it should be considered within the context of the evidence and not extrapolated beyond. Whether an intended optimum concerns the average response, reproducibility, or a future outcome can influence the decision on how to proceed. In this case, without specifying particular reaction settings, the focus is on understanding the relationship between the design, fitted model, and the scope of the recommendation, ensuring that the optimization does not lead to decisions based on conditions outside the studied region.

Selecting a condition changes its evidential status

When selecting a condition, its evidential status changes fundamentally. The condition chosen is no longer simply a representative of its expected performance. It has become an outcome of a selection process that introduces uncertainty. To address this, Benjamini and Hochberg (1995) outlined a multiplicity correction for defined families of tests. However, the selection uncertainty in our context extends beyond just attaching a multiplicity procedure to every coefficient. It encompasses the entire process of choosing one condition over others, making it crucial to acknowledge the selection itself rather than presenting the selected result as if no selection had taken place.

Sequential decisions can be sensible if documented, but they must not obscure the fact that the optimization claim rests on the outcomes of these selections. By preserving unfavorable candidate conditions, we provide a clear basis for the optimization claim without presenting a selected result as though the selection had never occurred. This approach ensures that the robustness and reproducibility of the optimization are not overlooked.

The marginal effect depends on a population

A marginal condition effect averages over a specific group or set of groups, rather than across the entire population. In an unbalanced design, the relevant context distribution might differ from the raw distribution of recorded rows. For example, if we consider a generic measurement and abstract settings, the context distribution could vary depending on the conditions under which data were collected.

The multilevel perspective in Gelman and Hill (2006) highlights this point, emphasizing that the reported contrast should be interpreted in relation to the observed groups or an equally weighted set of groups. It is crucial to distinguish between these interpretations because they can answer different scientific questions. A condition may be favorable in one context but less so in another, yet a single overall contrast obscures these differences.

To address this issue, a future assessment should preserve conditional results and clearly state the weights used for the principal summary. Additionally, it should specify which context combinations the design actually supports. By doing so, we connect an average reaction-condition recommendation with its intended population rather than allowing the population to be inferred from a coefficient label. This approach ensures that our conclusions are more robust and aligned with the specific conditions of the new batch.

The experimental unit determines the comparison

In the context of an experimental unit, the comparison between conditions should not be conflated with the variability inherent within that unit. When measurements are taken from a single experimental unit, they do not provide independent replication necessary for assessing batch-to-batch variation. This is a critical point when the goal is to draw inferences about a new batch rather than just the unit itself. According to Laird and Ware (1982), shared variation among units is relevant but must be accounted for within the design framework, not post hoc. Therefore, if conclusions are intended for a new batch, the precision observed within that batch cannot substitute for evidence of variability across different batches. To address this issue, a clear design description should distinguish between the average response and the variation specific to the target population. This ensures that measurement tables remain informative without overstating the number of independent replicates.

What the observation represents

In the hypothetical scenario of optimizing a measurement, one entry in a chart might represent an individual observation, capturing the unique variability at any given time. At a second level of aggregation, such as summarizing multiple observations into a batch, this unit's representation changes. When aggregating, the focus shifts from the variability within a single instance to the consistency or variation across the entire batch. This transition highlights that while an observed condition may have the highest estimated mean in a specific instance, it might not be the most dependable choice when considering the variability relevant to a new batch. In other words, the reliability of a condition as optimal depends on the context-whether we are looking at individual instances or aggregating them into batches.

Sorting records into comparable groups

When sorting records into comparable groups, it is plausible that including certain observations can alter the appearance of an observed association. For instance, consider a generic measurement, such as yield, from different batches. If we group these batches together based on some arbitrary criterion, the mean yield for the entire group may appear higher or lower than the average yield of individual batches. This grouping could mask or highlight variation that is specific to certain batches, thus changing our perception of the overall trend.

In contrast, if we sort the batches into groups based on a more meaningful criterion-such as batch composition, temperature, or time-this can provide a clearer picture of how variability affects the measurement. Each group will likely have its own mean yield, and comparing these means across groups can reveal patterns that are not apparent when all data are lumped together. This method allows us to distinguish between average responses and variation relevant to each new batch, as the article suggests.

By grouping the batches appropriately, we maintain a focus on the variability that is pertinent to the specific conditions under which we now wish to apply our optimization strategy. This approach enhances our understanding of how different factors influence the outcome and aids in selecting an optimal condition that is robust across all relevant batches.

Conclusion

One consequence of this transition is that the reliability of a condition as optimal can vary depending on the context-whether we are looking at individual instances or aggregating them into batches. Another open question is how to determine the most meaningful criterion for grouping batches, as this can significantly influence the perceived variability and the optimal condition.

References

Bates, D., Mächler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67(1)*. https://doi.org/10.18637/jss.v067.i01

Box, G. E. P., Behnken, D. W. (1960). Some New Three Level Designs for the Study of Quantitative Variables. *Technometrics, 2(4), 455-475*. https://doi.org/10.1080/00401706.1960.10489912

Benjamini, Y., Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 57(1), 289-300*. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Gelman, A., Hill, J. (2006). Data Analysis Using Regression and Multilevel/Hierarchical Models. . https://doi.org/10.1017/cbo9780511790942

Zhao, R., Zhang, J., Wang, X., et al. (2023). Research on optimal design of reaction conditions based on mixed statistical model. *Second International Conference on Applied Statistics, Computational Mathematics, and Software Engineering (ASCMSE 2023)*. doi:10.1117/12.2692066.

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

Satterthwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin, 2(6), 110*. https://doi.org/10.2307/3002019

Laird, N. M., Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics, 38(4), 963*. https://doi.org/10.2307/2529876

Kenward, M. G., Roger, J. H. (1997). Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. *Biometrics, 53(3), 983*. https://doi.org/10.2307/2533558

Box, G. E. P., Wilson, K. B. (1951). On the Experimental Attainment of Optimum Conditions. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 13(1), 1-38*. https://doi.org/10.1111/j.2517-6161.1951.tb00067.x