Heterodox Research Society

Something Hidden in the Scientific Process
Procedural Observability in a Simple Gravimetric Workflow

Zachary Lubick, Alexander Vawter

Abstract

There is something hidden in the scientific process: a layer of execution that lies between protocol and outcome, shaping how experiments unfold, but it is rarely captured in a structured or analyzable form. While protocols define intended steps, the actual sequence of actions - handling, timing, and interaction with materials - remains largely unobserved and unstructured in standard laboratory practice. This missing layer is especially relevant to problems of reproducibility, where experimental variability is often discussed in terms of technical error or poor reporting, while the actual process of execution remains only weakly documented and difficult to compare across runs.

In this study, we demonstrate procedural observability using a simple gravimetric workflow. Ten nominally identical runs were performed while capturing event-level procedural traces via *Experiment Engine*, a computer vision system that records timestamped object interactions and laboratory events. Despite consistent materials and intended procedure, all runs exhibited systematic mass loss, with structured variation emerging across trials.

Procedural traces revealed substantial differences in execution, with total recorded events ranging from 121 to 232. Patterns such as systematic underfilling of final aliquots and differences in interaction structure show that variability arises through the sequence of execution rather than from isolated error.

These findings do not establish causal relationships between specific procedural metrics and outcome. Instead, they demonstrate that experimental execution constitutes an observable layer that can be structured and analyzed. Making this layer visible enables deviations to be interpreted as consequences of how experiments are performed, expanding what can be measured and understood in scientific practice.

Introduction

Scientific experiments are typically understood through protocols and results, where a methodological pathway is defined, executed, and evaluated by its outcome under the assumption that the procedure is stable and reproducible. When results differ, measurement error, environmental noise, or uncontrolled variables are often invoked as explanations. Procedural execution is also understood to play a role, but in practice, it is usually accessible only through memory or sparse notes, and is rarely captured in a form that allows for systematic examination.

In reality, however, protocols are rarely as reproducible as they seem. The fine-grained procedural actions - handling, transferring, pausing, correcting - are shaped by timing, context, and human interaction, forming a dense and variable layer of activity that is only minimally captured in typical laboratory reports. This layer can influence experimental outcomes, yet remains difficult to compare across runs. This becomes especially relevant in the context of reproducibility, where scientific work is widely recognized as a messy and variable process in practice, but is typically communicated in a far cleaner and more compressed form. As a result, when deviations occur, the procedural pathways that produced them remain largely inaccessible.

This gap becomes particularly visible in simple gravimetric workflows, where mass conservation provides a clear and quantitative reference. Under nominally identical conditions, systems are expected to produce identical outcomes, making even small deviations measurable and comparable. In this study, we examine a minimal example: a micro-batch aqueous formulation of caffeine in deionized water, prepared and aliquoted across ten independent runs using the same materials, equipment, and intended procedure.

Despite the experiment's simplicity, all runs exhibited measurable mass loss when compared to their theoretical batch mass, with deviations ranging from approximately 0.2g to 0.4g. These losses were not uniformly distributed, as within each run the first four aliquots were consistently near the target mass while the final aliquot was systematically underfilled due to the residual effects of prior processing steps. This pattern suggests that variability is not only present in outcomes, but emerges through the sequence of experimental steps themselves as the procedure unfolds.

Importantly, the presence of deviation in such systems is not unexpected. Manual liquid handling, transfer loss, and surface adhesion are well-understood contributors to mass imbalance in small-scale workflows. The existence of error is therefore not the central problem. Rather, the problem is that these deviations are typically uninterpretable. Standard laboratory records do not preserve the sequence of actions that produced them, making it difficult to distinguish between different procedural pathways that lead to similar outcomes.

To investigate this further, we employed a system for capturing event-level traces during experimentation, where computer vision was used to record timestamps associated with object interactions and laboratory events. This produced a structured representation of how each run unfolded, allowing for comparisons between the procedural structure of trials and enabling deviations to be examined in the context of their underlying execution rather than only their final measurements.

We refer to the ability to capture, structure, and analyze the sequence of events that takes place during experimental work as procedural observability. This study presents proof of existence for this concept, showing that even in a simple gravimetric system, nominally identical protocols can produce systematically biased outcomes based on how the experiments are actually carried out. By making this

layer observable, results can be interpreted in the context of execution rather than only final measurements.

Experimental System and Workflow

To examine variability under nominally identical conditions, this study focused on a simple gravimetric workflow. Each run consisted of an aqueous caffeine solution prepared by combining approximately 1g of anhydrous caffeine with approximately 49g of deionized water, yielding a nominal batch mass of 50g and a target concentration of 2% w/w. A precision balance with 1mg resolution was utilized for all measurements. The solutions were prepared in a glass beaker by sequentially weighing water and caffeine then manually mixing until the powder was dissolved.

Rather than enforcing exact target masses, the actual measured amounts of water and caffeine were recorded and used to define the theoretical batch mass for each run. Across the ten runs, total batch masses ranged from 49.916g to 50.167g, reflecting typical variability around the nominal target.

The prepared solution was then aliquoted sequentially into five weigh boats, targeting 10g per aliquot. The first four aliquots were distributed near the target mass, while the fifth aliquot represented the remaining solution in the beaker. As a result, the final aliquot reflected the cumulative effects of prior handling steps. Verification of each run was performed by summing the recovered masses across all aliquots and comparing this value to the theoretical batch mass. This mass balance provided a quantitative framework for assessing deviation.

All ten runs were performed using the same materials, equipment, and written protocol. No disruptions were intentionally introduced. The observed differences were solely due to the variations in procedural execution.

Procedural Trace Capture

Experiment Engine, a computer vision-based system, was used to generate event-level procedural traces during each run. This system employs a fixed overhead camera to continuously record the benchtop, allowing for object interactions and laboratory events to be inferred as they occurred.

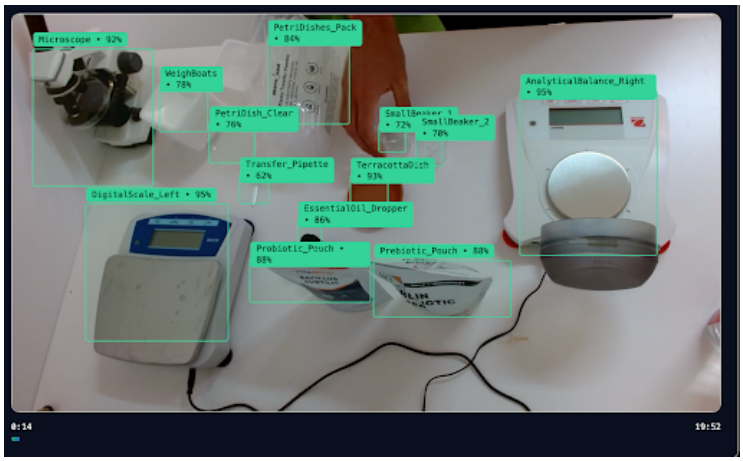

Figure 1. Example frame from *Experiment Engine* vision system showing real-time object detection and region-of-interest (ROI) labeling during experimental execution. Bounding boxes identify laboratory objects and tools, with associated confidence scores, enabling continuous tracking of object interactions and contributing to the generation of event-level procedural traces.

Events with timestamps were extracted from the system's visual input based on activity such as object movement, batch transfer and dispensing, and aliquot handling. An event is defined as a single Lab Log entry generated by the system (described below), representing a vision-to-LLM prediction of observed activity within a given temporal window. Each event corresponds to a structured record in the Lab Log output, containing a timestamped natural language description of inferred activity occurring within the experimental workspace.

At a practical level, the system generates two parallel outputs: Lab Logs and Raw Data. Lab Logs are vision-to-LLM predictions generated on a scheduled frequency of approximately 20 seconds, with slight processing latency. Each prediction window may produce multiple logs if more than one action is inferred to have occurred within that sequence. These logs take the form of full sentence descriptions of observed activity at a given timestep. Lab Log predictions are generated using a combination of contextual inputs, including the experiment's provided name and description, associated protocols, detected objects within the workspace, and the live video feed of operator interactions. This allows the system to generate predictions with awareness of the intended experimental process as it unfolds.

Raw Data, in contrast, consists of lower-level ROI interaction signals that describe when and where engagement occurs within the frame. While these interaction events do not directly translate into Lab Logs, they provide an additional layer of real-time context that informs the vision-to-LLM prediction process. While Raw Data may provide supporting context, it was not used in the procedural trace analyses presented in this study; Lab Logs were the primary object of analysis.

Examples of both outputs are shown below:

Example Lab Log:
hash: e8446a6b | description: "@researcher1 poured DI water from DIWaterBottle into Beaker1." | id: exp-1772143309191_1543849_1573800-action-5 | ts: 17:01:49

Example Raw Data Item:
ts: 19:25:17 | interaction_id: act_5a118d7a | roi: weighingboat3 | event: roi-interaction | conf_interval: 0.25 | exp: exp_17721423... | lab: Heterodox | user: x

Figure 2. Example outputs from *Experiment Engine* showing both Lab Logs (top) and Raw Data (bottom). Lab Logs consist of natural language descriptions of inferred laboratory actions at specific timestamps, and Raw Data captures high frequency ROI interaction events with associated metadata and confidence intervals.

As a prototype system, the outputs are not free from error. Vision-based detection and LLM-generated descriptions may occasionally misclassify objects or events, and confidence estimates associated with ROI interactions are approximate. However, the system provides a consistent representation of the structure, timing, and density of procedural execution, which is the focus of this study.

In this work, procedural traces are used descriptively rather than as inputs to a predictive model. Their role is to make experimental execution analyzable, enabling differences in sequencing, interaction frequency, and handling structure to be examined across runs.

Outcome Variability and Procedural Trace Inspection

Ten experimental runs were performed using the workflow described above. While each run was carried out by the same researcher and followed the same nominal procedure under similar conditions, measurable mass loss was observed in all cases.

The theoretical batch mass for each run was defined as the sum of the measured water and caffeine inputs. The total recovered mass was calculated by summing the five aliquots, and deviation was determined by comparing recovered mass to the theoretical value. Across the ten runs, recovery ranged from 99.23% to 99.59%, corresponding to losses of 0.206g to 0.387g.

In addition to total mass loss, variability was observed in the distribution of mass across aliquots within each run. The first four aliquots were generally near the target of 10g, while the fifth aliquot consistently reflected the cumulative effects of prior transfers.

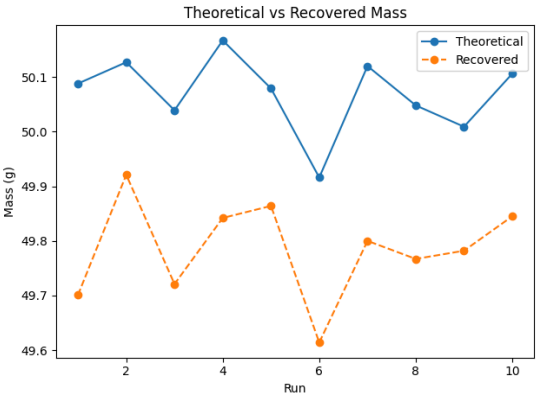

Figure 3. Comparison of theoretical batch mass and recovered mass across ten the ten runs

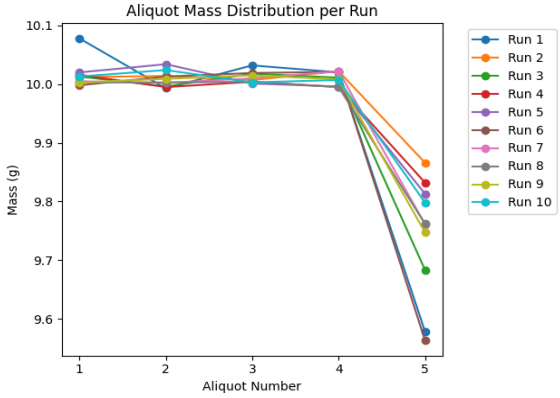

Figure 4. Aliquot mass distribution across the ten runs. The first four aliquots remained near the 10g target, while the fifth aliquot was consistently underfilled.

While the pattern of deviation is clear at the level of final measurements, the underlying cause is not immediately evident. To investigate how these deviations arise during execution, procedural traces were examined for individual runs.

Two runs with contrasting levels of deviation were analyzed using their corresponding Lab Logs. The structure of execution becomes more apparent when examining short sequences of entries. Selected excerpts from the corresponding Lab Logs are shown below (ordered and lightly formatted for readability):

Run 1 (Higher Deviation)

21:23:12 - @researcher used the Stirring Rod to stir the clear liquid contents
21:23:48 - @researcher continued to stir the clear liquid contents
21:24:20 - @researcher placed the Stirring Rod into a container

21:25:20 - @researcher picked up the Micropipette
21:25:43 - @researcher aspirated liquid from Beaker_2 using the Micropipette
21:25:43 - @researcher dispensed liquid into a Weighing Boat
21:26:58 - @researcher dispensed liquid into Weighing Boat '1'

Run 2 (Lower Deviation)

22:03:44 - @researcher used StirringRod1 to stir the contents
22:03:44 - @researcher placed StirringRod1 back onto the lab bench

22:04:41 - @researcher attached a PipetteTip to the Micropipette
22:04:41 - @researcher aspirated liquid from DIWaterBottle
22:04:41 - @researcher dispensed liquid into WeighingBoat1

22:05:32 - @researcher aspirated liquid from DIWaterBottle
22:05:32 - @researcher dispensed liquid into WeighingBoat1

The traces show differences in how mixing and transfer operations were structured. In the higher-deviation run, an extended mixing segment is followed by dense pipette activity, indicating close coupling between mixing and aliquoting. In the lower-deviation run, mixing appears as a discrete event followed by a later transition to pipette operations, indicating greater temporal separation between these steps.

These differences correspond to distinct interaction patterns between the solution and laboratory tools. Inspection of the experimental video provided additional context: in the higher-deviation run, the stir rod was used in close proximity to transfer operations, resulting in repeated contact during aliquoting. In the lower-deviation run, stirring was more isolated from the transfer phase, reducing such interactions.

The procedural trace does not directly measure material loss, but it captures the structure of interactions that give rise to it. The correspondence between trace structure and outcome shows that variability emerges from how steps are executed and integrated within the workflow.

Procedural Differences Across Runs

Despite following the same nominal protocol for each run, analyzing the procedural traces as documented by *Experiment Engine* revealed substantial differences in how the experiments were carried out. These differences are directly observable within the Lab Logs, and can also be summarized through aggregate measures such as event counts and procedural density. While the same protocol was followed across all ten runs, the number of recorded events ranged from 121 to 232, leading to variations in procedural density. Some runs showed a linear workflow from batch formulation to aliquoting, while others had increased numbers of dispensing, repositioning, and intermediate handling steps. These differences are summarized in Table 1.

The following table summarizes total recorded events, experimental duration, and calculated procedural density across all runs.

Run	Total Events	Duration (min)	Procedural Density (events per minute)
1	189	12.89	**14.67**
2	123	11.09	**11.10**
3	232	15.40	**15.07**
4	178	15.15	**11.75**
5	157	11.62	**13.51**
6	168	13.70	**12.26**
7	184	13.39	**13.75**
8	152	11.32	**13.43**
9	142	12.17	**11.67**
10	121	11.89	**10.18**

Table 1. Summary of procedural density across experimental runs, including total recorded events, experiment duration, and calculated event rate.

The values in Table 1 illustrate that procedural execution varied meaningfully across runs, despite identical protocols. Procedural density ranged from 10.18 to 15.07 events per minute, reflecting differences in how actions were distributed over time. Higher-density runs indicate periods of more frequent interaction with materials and equipment, which may arise from factors such as corrective

actions, repositioning, or repeated handling within short time windows. Lower-density runs, in contrast, reflect more temporally spaced execution, where steps were carried out in a more linear and segmented manner.

These differences do not indicate changes in the intended procedure, but rather how the procedure was realized in practice. The table therefore provides a coarse summary of executional variation, highlighting that even under controlled conditions, the structure and pacing of experimental activity can differ substantially across runs.

While aggregate measures such as total events and procedural density characterize overall differences in execution, more specific relationships can be examined by focusing on individual categories of actions. One such category is pipette-related events, which directly correspond to liquid transfer operations and therefore represent a plausible contributor to mass deviation.

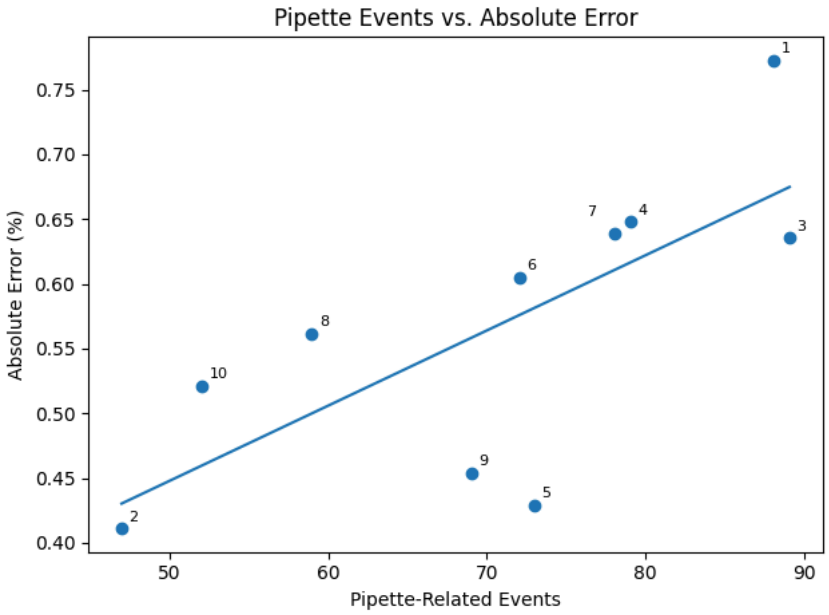

Figure 5. Relationship between pipette-related procedural events and absolute gravimetric error across experimental runs.

The relationship between pipette-related events and absolute mass deviation is shown in Figure 5. A positive trend is observed, where runs with higher numbers of pipette events tend to exhibit greater mass loss. For example, Run 1 shows one of the highest counts of pipette events and also exhibits the highest observed deviation. In addition, it has the second highest total event count, indicating a generally dense execution pattern.

However, this relationship is not strictly monotonic. Run 3, which has the highest number of pipette events, does not exhibit the highest deviation. This suggests that pipette activity alone does not fully account for the observed variability. Instead, the relationship between execution and outcome appears to be influenced by multiple interacting factors.

This interpretation is consistent with the procedural trace inspection presented earlier. In the case of Run 1, deviation was associated not only with pipette activity, but also with how mixing and transfer operations were structured, including repeated interaction between the solution and the stir rod during aliquoting. These additional interactions introduce alternative pathways for material loss that are not captured by pipette counts alone.

Taken together, these observations suggest that experimental deviation is not determined by a single category of action, but emerges from the combined structure of execution. Aggregate measures such as pipette event counts can highlight potential relationships, but the underlying procedural record is required to interpret how those relationships arise in practice.

Event Category	Run 10 (Low Density)	Run 6 (Intermediate Density)	Run 3 (High Density)
Pipette Events	43.8%	43.5%	38.8%
Container Handling	19.8%	18.5%	26.7%
Balance Interaction	32.2%	31.0%	23.3%
Documentation	3.3%	3.6%	2.2%
Material Handling	0.8%	1.2%	4.3%
Other /Unclassified	0.0%	2.4%	4.7%

Table 2. Refined semantic action distributions across low-, intermediate-, and high-density runs. Categories were consolidated to emphasize dominant behavioral modes. Higher density execution is characterized by increased container and material handling, while lower density runs exhibit greater emphasis on balance interactions, indicating a shift in execution structure across runs.

In addition to overall activity levels, the composition of actions within each run provides further insight into how execution differed. The distribution of event categories across low, intermediate, and high-density runs is shown in Table 2.

While pipette-related events constitute the largest proportion across all runs, the relative contribution of other categories shifts with overall execution structure. In the high-density run (Run 3), a greater proportion of container and material handling events is observed, indicating increased interaction with physical components of the workflow beyond liquid transfer alone. In contrast, the low-density run (Run 10) shows a higher relative proportion of balance interaction

events and a more constrained distribution of action types, consistent with a more linear execution pattern.

These differences highlight that procedural variability is not only reflected in the number of actions performed, but also in the types of actions that make up the workflow. Variation in execution therefore arises from both the frequency and the composition of interactions, with different runs exhibiting distinct procedural profiles.

This level of categorization is enabled by the structure of the procedural trace itself. Because Lab Logs capture individual actions as discrete, timestamped events, they can be grouped into higher-level categories such as pipetting, container handling, and balance interaction. This allows execution to be analyzed not only as a sequence, but also as a distribution of activity types within a run.

While the specific categories in this study reflect the simplicity of the gravimetric workflow, the underlying approach is not limited to this setting. In more complex experimental systems, similar categorization could be applied to domain-specific actions, such as sample preparation, incubation, imaging, or assay timing in biological workflows. The ability to capture and organize these events suggests a general framework for analyzing experimental execution across different disciplines, where both the sequence and composition of actions contribute to observed outcomes.

Implications for Experimental Reproducibility and Lab Documentation

Concerns regarding reproducibility in scientific research have been well documented across disciplines. A 2016 survey published in *Nature* reported that a majority of researchers had experienced difficulty reproducing their own experiments and those of others, highlighting systemic challenges in experimental reliability and interpretation. While these issues are often attributed to factors such as methodological differences, insufficient reporting, or statistical practices, the role of procedural execution remains less directly observable.

The results of this study suggest that procedural execution constitutes a measurable component of experimental variability, even in simple systems. By capturing event-level traces of experimental activity, procedural observability introduces a structured record of execution that can complement traditional laboratory documentation. This enables differences in outcomes to be examined in relation to how experiments were performed, rather than solely in terms of inputs and results.

In this context, procedural traces function as a form of experimental record that extends beyond protocol description. They allow researchers to inspect the sequences, timing, and interaction patterns that occur during experimentation, providing a basis for interpreting variability and diagnosing

deviations. This type of record may be particularly useful in cases where results fall within acceptable ranges but exhibit systematic differences that would otherwise remain unexplained.

Procedural observability does not eliminate variability or guarantee reproducibility. Instead, it provides a framework for making experimental execution visible and analyzable. By expanding what is recorded during experimentation, it enables a more complete account of how results are produced, supporting improved interpretation, comparison across runs, and refinement of experimental practice.

Conclusion

It was demonstrated that even a simple procedure, like a gravimetric workflow, contains a structured layer of procedural variability that is not captured by conventional laboratory records. Across ten nominally identical runs, systematic mass loss in aliquot distribution was observed. By capturing event-level procedural traces, the outcome differences could be interpreted in terms of how the experiments were actually carried out. Variability was not only present, but sometimes structured, and emerged through the sequence of events that unfolded during execution.

More broadly, the concept of procedural observability provides a framework for making this hidden layer visible and analyzable. Rather than treating deviations as unexplained errors, they can be examined in the context of how experimental procedures are actually performed. This approach preserves variability while making it interpretable as a consequence of execution.

The central contribution of this work is the demonstration that this layer exists and can be made observable. Bringing it into view expands the scope of what constitutes an experiment, from a sequence of prescribed steps to a process whose structure can be measured, analyzed, and understood.