capstract:2611.00073v1 [stat.ME] 20 Nov 2026

Recursive Review of Meeting Substitution: Partial Identification and Queue Length

Iona F. Vardale Nadia V. Ellery
Department of Statistics, Westhaven University, Bristol, United Kingdom

November 20, 2026

Abstract

A meeting-review committee can decide whether a proposed meeting should be replaced by email. When the committee convenes a review meeting, that meeting may itself be subject to review. We define the replaceable fraction of successful review meetings using potential outcomes for synchronous and email review, derive sharp bounds conditional on review categories, and construct a simultaneous finite-sample confidence envelope. In a simulated comparison of 2,000 review items, category-specific estimated bounds are [0.747, 0.840], the 95% confidence envelope is [0.654, 0.943], and the generating value is 0.797. We then analyze a recursive review rule in which each scheduled meeting independently generates a preliminary review meeting with probability rho. At rho = 0.8, four review invitations are generated on average before the earliest review begins; at rho = 1 the rule has no first review. An oracle substitution policy reduces the expected number of review invitations, but its effect depends on the joint outcome distribution that the trial does not identify. The numerical example is simulated and does not estimate the prevalence of the practice in workplaces.

Keywords: partial identification, potential outcomes, randomized experiments, meetings

1 Introduction

An organization may appoint a committee to review a scheduled meeting and decide whether its business could instead be conducted by email. If the review itself is scheduled as a meeting, it too is eligible for review. The decision to convene a meeting can therefore create another meeting whose necessity has yet to be assessed.

This paper considers a registry of proposed meeting-review decisions. A review can be conducted either synchronously or by email, and a decision is successful if a recommendation is recorded within a specified period. The question is how often the synchronous reviews that would succeed could have produced a successful recommendation by email. Our target is a counterfactual fraction of review meetings, not the fraction of the original meetings that should be canceled.

Research on meeting load has examined employee well-being (Luong and Rogelberg, 2005). A randomized comparison of the two review methods can estimate their rates of timely decisions, but a review item can be observed under only one method (Rubin, 1974). Whether both methods would have succeeded on the same item is not directly observed. We use bounds on joint potential outcomes (Tian and Pearl, 2000) to describe the resulting uncertainty and apply them to a simulated comparison. We then study the consequences of requiring review meetings to undergo the same review procedure before they begin. The numerical example is simulated throughout.

2 Definition and study design

An item is a decision on whether a proposed parent meeting should be replaced by an email. It is eligible if the proposed meeting has a written purpose, named participants and a decision owner for its review. The interventions are alternative methods of conducting that review. The parent meeting may concern ordinary business or may itself be a review meeting. The interventions are (M) a scheduled synchronous discussion followed by a written record and (E) an email to the same recipients requesting a written decision, with no meeting during the one-hour assessment period. Both start at time zero. A success is a recommendation to retain or replace the parent meeting, recorded within 60 minutes and accompanied by a reason satisfying prespecified criteria. The same criteria apply in both arms. The 60-minute horizon is a design choice, not a claim that all email is answered within an hour.

Let Y_M and Y_E be the binary outcomes for the two interventions. The estimand is

$$\begin{aligned} \theta &= \Pr(Y_E = 1 \mid Y_M = 1) = \frac{q}{p_M}, \\ p_M &= \Pr(Y_M = 1), \quad p_E = \Pr(Y_E = 1), \quad q = \Pr(Y_M = 1, Y_E = 1). \end{aligned} \tag{1}$$

with $p_M > 0$. Replaceability refers here to achieving the same *standard of completion*, not necessarily the same wording or the same recommendation. The denominator excludes review decisions for which the synchronous review itself would not have succeeded. We do not call such a meeting replaceable even if an email would have succeeded; that is a distinct class of intervention response. Nor does completion establish that either recommendation is correct, or that email is preferable after accounting for effort, quality beyond the rubric, and the consequences of an incorrect recommendation.

In a proposed trial, eligible review items are classified before assignment and then allocated at random within category to M or E. Only the assigned outcome is used. The potential outcomes require well-defined versions of the interventions, consistency of the recorded outcome with the assigned method, and no interference between items. In practice, repeated personnel, knowledge carried from one review item to another, and messages sent outside the assigned channel can violate the last two conditions. Randomizing at the team level and allowing adequate separation between items could reduce interference, but would require uncertainty calculations that account for clustering. The calculations below instead assume independent items, complete adherence and complete outcome ascertainment.

3 Identification

Randomization identifies p_M and p_E but does not identify q. Set $\pi_{ij} = \Pr(Y_M = i, Y_E = j)$. For any fixed marginals, $\pi_{11} = q$, $\pi_{10} = p_M - q$, $\pi_{01} = p_E - q$, and $\pi_{00} = 1 - p_M - p_E + q$. Nonnegativity of these four probabilities yields

$$\frac{\max(0, p_M + p_E - 1)}{p_M} \le \theta \le \frac{\min(p_M, p_E)}{p_M}. \tag{2}$$

These limits are sharp: every value between them produces a valid four-cell distribution, and all such distributions give the same marginal results in a randomized trial. The trial identifies the mean difference $p_E - p_M$ even when it cannot identify the fractions harmed and helped by re-

placement, π_{10} and π_{01}. Those two fractions can vary together while their difference remains fixed.

Suppose a prespecified review category X has levels x, with target-population weights w_x. Write $m_x = \Pr(Y_M = 1 \mid X = x)$ and $e_x = \Pr(Y_E = 1 \mid X = x)$. Conditioning gives the sharp bounds

$$\frac{\sum_x w_x \max(0, m_x + e_x - 1)}{\sum_x w_x m_x} \le \theta \le \frac{\sum_x w_x \min(m_x, e_x)}{\sum_x w_x m_x}. \tag{3}$$

The bounds are sharp because the joint distribution can attain either endpoint separately within each category. They are at least as tight as Eq. (2) with the same standardized marginals. The category must be recorded before the intervention; classifying a meeting by what happened during it would change the conditioning variable under email. The weights must refer to the population for which the question is asked. In the example below they are known by construction, rather than estimated from another sample.

A point estimate is available under an additional cross-world assumption. Conditional independence, $Y_M \perp Y_E \mid X$, gives $q = \sum_x w_x m_x e_x$. The trial cannot verify this assumption because no item reveals both outcomes. In the other direction, assuming that every meeting success would also have succeeded by email sets $\theta = 1$, but is compatible with the observed marginals only when $p_E \ge p_M$. Assuming instead that email never succeeds when a meeting would fail sets $q = p_E$ and gives $\theta = p_E / p_M$, provided $p_E \le p_M$. Neither monotonicity assumption follows from random assignment. In particular, the first would be contradicted by the population margins in our example, while the second would be contradicted by its routine-category margins.

4 Sampling uncertainty

The endpoints in Eq. (3) are population quantities. Substitution of sample proportions estimates an identified set; the resulting interval is not a confidence interval. We report a conservative finite-sample confidence envelope for the whole set (Imbens and Manski, 2004). For each of the $2K$ independently sampled category-by-arm means, construct a binomial Chernoff interval. If $\hat{p}$ is a sample proportion from n independent Bernoulli observations, retain values p satisfying

$$n\, D(\hat{p} \| p) \le \log(4K/\alpha), \quad D(a\|b) = a \log \frac{a}{b} + (1-a) \log \frac{1-a}{1-b}, \tag{4}$$

with the usual limiting conventions at zero and one. The two one-sided binomial tail bounds and a union bound give simultaneous coverage of all $2K$ means of at least $1 - \alpha$ (Hoeffding, 1963). No independence between arms is needed for the union bound; independent Bernoulli items are needed within each arm and category. We use $\alpha = 0.05$.

We minimize the lower endpoint and maximize the upper endpoint of Eq. (3) over this rectangle of marginal rates. The resulting envelope covers the entire population identified set whenever the rectangle covers the true rates. It can be wider than a confidence interval designed to cover only the unknown value of θ, and it is generally conservative. It does not account for estimated category weights or clustered items. The endpoint optimization used here is described in Appendix A.

5 A simulated comparison

We generated 2,000 independent review items in three prespecified categories. For each category, an equal number was assigned to each intervention. The target weights and both potential-outcome distributions were specified before simulation, with no workplace observations used. Each pair (Y_M, Y_E) was drawn independently from its four-cell distribution; assignments were then randomly balanced within category. Only the assigned outcome enters the analysis. A further 2,000 simulated trials assess coverage of the interval procedure. The random seed and generator are given in Appendix B.

Table 1 gives the generating margins and joint success probabilities. Routine reviews are relatively likely to succeed under either method, whereas escalation reviews have lower success probabilities. The categories are labels for the simulation, not empirically estimated classes. The joint probability is shown only because it is known in the simulation; an actual trial would not observe it.

Table 1: Generating distribution for the simulated trial. q_x is the joint success probability, unavailable from observed trial outcomes.

Category	Weight w_x	m_x	e_x	q_x	Items per arm
Routine	0.50	0.90	0.95	0.875	500
Reconciliation	0.25	0.96	0.72	0.700	250
Escalation	0.25	0.44	0.18	0.100	250
Overall	1.00	0.80	0.70	0.638	1,000

The true overall joint probability is 0.6375, giving $\theta = \mathbf{0.7969}$. Knowing only the population margins leaves the interval [0.625, 0.875]; using the category-specific margins narrows it to [0.7438, 0.8438]. The true value lies in both. At the lower category-specific endpoint, 59.5% of all items succeed under both methods; at the upper endpoint, 67.5% do. Every joint distribution in that range can reproduce the same category-specific trial outcomes.

The observed results from one simulated trial are in Table 2. The overall rates are 0.792 and 0.702, a difference of $\mathbf{-0.090}$ for email minus meeting. The plug-in pooled bounds are [0.624, 0.886], and the category-specific bounds are [0.747, 0.840]. Conditional independence gives 0.777, inside those bounds but below the generating value of 0.797. The 95% simultaneous confidence envelope for the category-specific set is [0.654, 0.943]. It includes both sampling and identification uncertainty; its width is not a standard error.

Table 2: Assigned outcomes in one simulated trial. Every item is counted in exactly one arm.

Category	Meeting success / assigned	Email success / assigned
Routine	442 / 500	479 / 500
Reconciliation	239 / 250	182 / 250
Escalation	111 / 250	41 / 250
Overall	792 / 1,000	702 / 1,000

Across 2,000 independently generated replications, the envelope covered the entire true identified set in 2,000 trials and covered the generating value of θ in 2,000 trials. Its mean width was 0.295. These counts illustrate conservatism in this design; they do not establish coverage in other designs. The maximum binomial Monte Carlo standard error for a coverage proportion based on 2,000 replications is 0.0112. The guarantee instead comes from the simultaneous concentration bounds and the assumed sampling design.

Figure 1 separates the effect of increasing the number of items from the width that remains when all category margins are known. The sample-size curves are design calculations evaluated at the generating margins, not intervals from additional simulated trials. Their residual gap is due to unobserved joint outcomes and does not vanish with more randomized items.

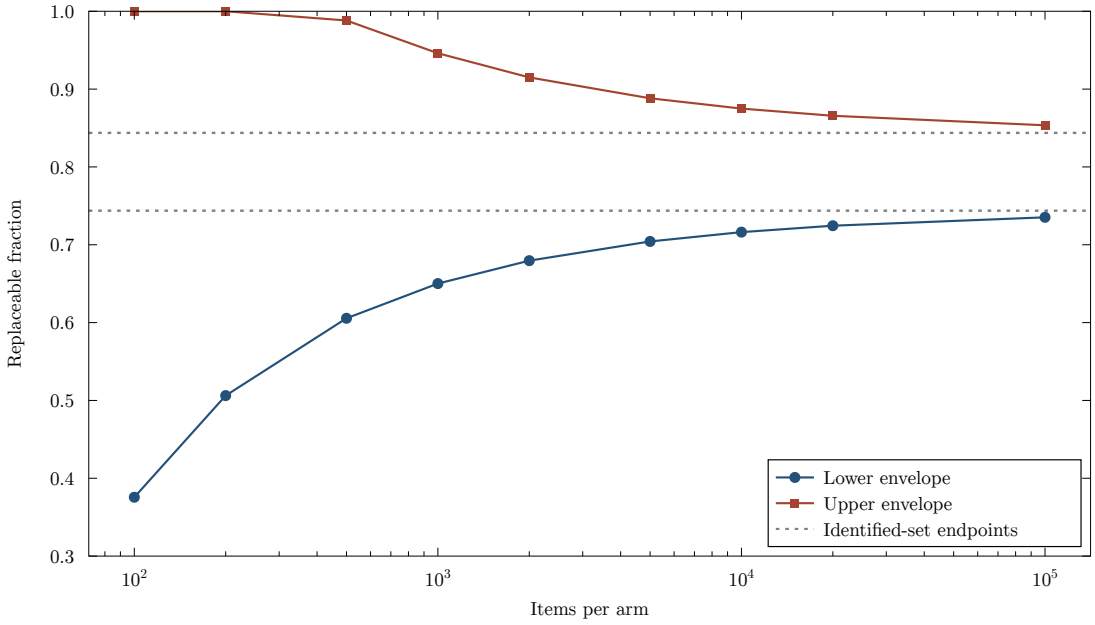


Figure 1: Design calculation for the simultaneous 95% confidence envelope as sample size increases. Equal arm allocation and category weights from Table 1 are held fixed. Curves use proportions equal to the generating margins; dotted horizontal lines are the population identified-set endpoints.

6 Implications for replacement decisions

A decision maker may care differently about an item that would succeed in a meeting and fail by email and an item with the reverse pattern. Let the loss from replacing all review meetings be

$$L = c_H \Pr(Y_M = 1, Y_E = 0) - c_B \Pr(Y_M = 0, Y_E = 1) = c_H p_M - c_B p_E + (c_B - c_H) q. \quad (5)$$

The first term penalizes a lost decision and the second credits a decision gained. For equal weights the joint probability cancels and $L = p_M - p_E$, so a randomized comparison identifies this loss. For unequal weights it generally does not. This loss function is an illustration of asymmetric valuation, not a measurement of the financial cost of either method.

Take $c_H = 0.4$ and $c_B = 1$ on a common arbitrary scale. With the generating margins, $L = -0.38 + 0.6q$. The category-specific bounds permit q from 0.595 to 0.675, so L ranges from

-0.023 to 0.025 per item. Both signs agree with the observed population marginals. The generating joint probability gives $L = 0.0025$. The same trial could therefore support either ranking under this specified asymmetric loss even though the mean success difference is fixed. Figure 2 shows where the ranking changes; the numbers do not prescribe an organization's values.

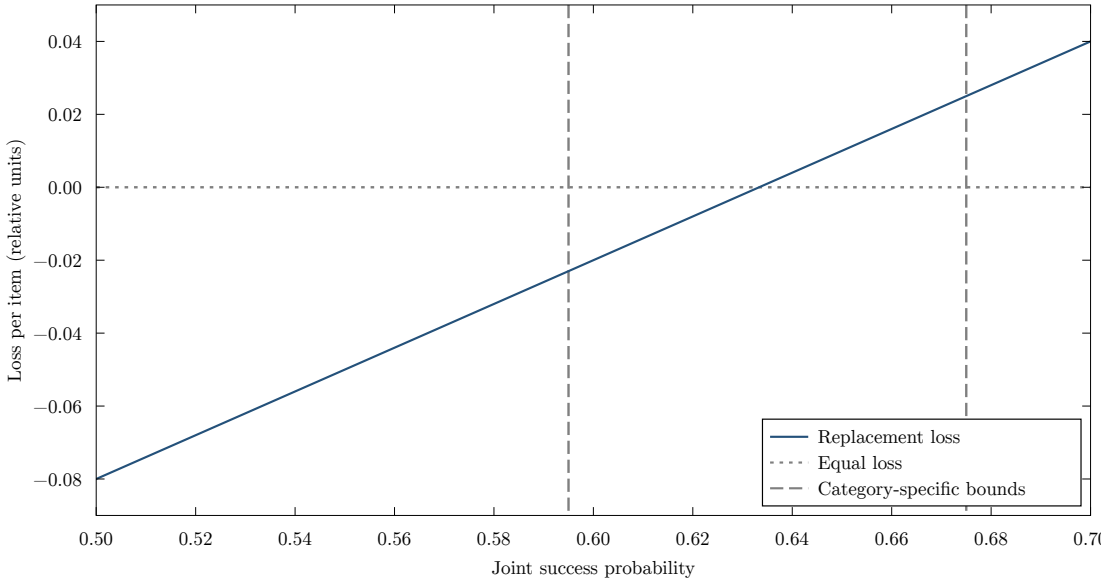


Figure 2: Loss per item from replacing review meetings by email for a specified asymmetric loss. The vertical lines delimit the population range of joint success allowed by category-specific margins; negative values favor email.

7 Recursive review of review meetings

Consider a rule under which a scheduled meeting generates a preliminary review meeting with probability ρ. The review meeting asks whether its parent meeting can be replaced by email. Because the review is itself a scheduled meeting, the same rule is applied to it, independently at each level. Invitations for all preliminary reviews are generated before the earliest review can begin. Let N be the number of review-meeting invitations generated for a single initial meeting. Under the independent rule,

$$\Pr(N = n) = (1 - \rho)\rho^n, \qquad \mathbb{E}N = \frac{\rho}{1 - \rho}, \qquad \operatorname{Var}(N) = \frac{\rho}{(1 - \rho)^2}, \tag{6}$$

for $n = 0, 1, \ldots$ and $0 \le \rho < 1$. Also $\Pr(N \ge k) = \rho^k$. These are invitations awaiting review, not necessarily meetings ultimately held: a later recommendation may cancel an earlier invitation. At $\rho = 1$ the chain is infinite. Since each review must be preceded by its own review, there is no first review to start. At $\rho = 0.8$, the expected number of review invitations is four and the probability of at least ten is 0.107. Table 3 shows the sensitivity to the rule's coverage.

Table 3: Invitations generated before the first review begins, per initial meeting. The mandatory-review rule is independently applied at each level.

Review probability ρ	Expected invitations	Probability of ten or more
0.25	0.333	0.000001
0.50	1.000	0.000977
0.80	4.000	0.107374
0.90	9.000	0.348678
0.99	99.000	0.904382

Suppose an oracle can decide, without holding an additional review, whether a proposed review meeting would succeed both as a meeting and as an email. It replaces exactly those review meetings by email. Review emails do not themselves enter the meeting-review rule. Assume the oracle's decisions are independent across levels, and each proposed review has joint success probability q from Eq. (1). The probability of another review-meeting invitation is then $a = \rho(1-q)$. If $a < 1$, the expected numbers of review-meeting invitations N_M and review emails N_E are

$$\mathbb{E}N_M = \frac{\rho(1-q)}{1-\rho(1-q)}, \qquad \mathbb{E}N_E = \frac{\rho q}{1-\rho(1-q)}. \tag{7}$$

For the second expression, an email is generated if the chain first produces any number of meeting reviews and then an email. A review email ends the chain. The calculation assumes that each level has the same mixture of categories and potential outcomes as the trial population. It does not require every review to recommend replacement of its parent. It counts invitations before recommendations are issued and does not model their subsequent execution.

At $\rho = 0.8$, the category-specific population bounds $0.595 \le q \le 0.675$ imply between 0.351 and 0.479 review-meeting invitations per initial meeting under the oracle rule, versus four without it. The corresponding range of review emails is 0.704 to 0.730. The generating value $q = 0.6375$ gives 0.408 meeting invitations and 0.718 emails. These are model calculations, not estimates of organizational savings; the oracle requires the unobserved joint outcomes and has no resource cost. Figure 3 shows how the expected number changes with the joint success probability and the review probability.

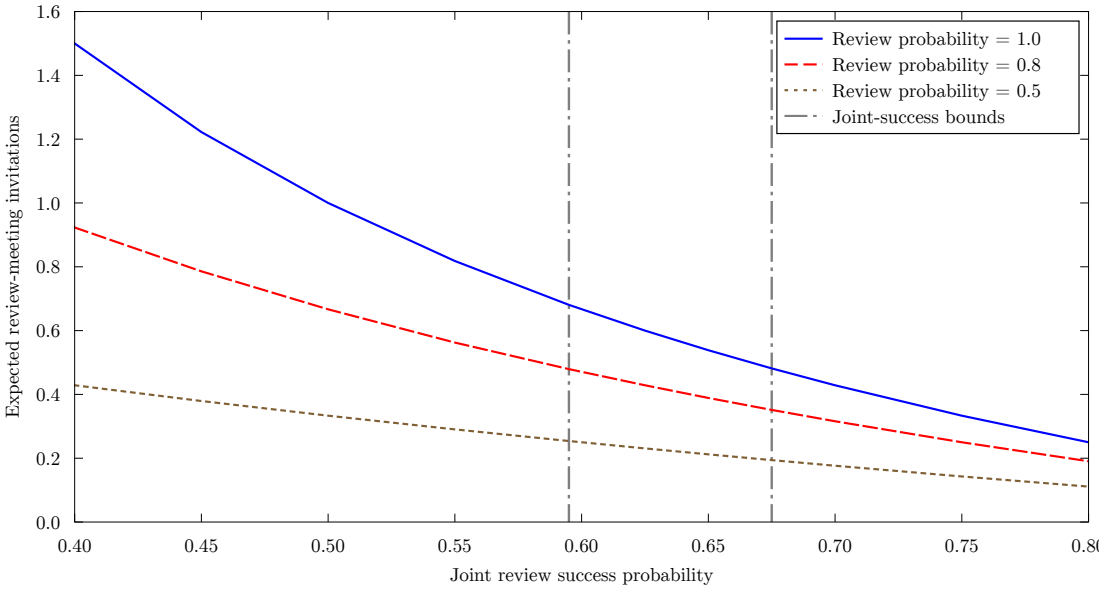


Figure 3: Expected review-meeting invitations under the oracle rule as a function of joint review success. Each curve uses a different probability that a scheduled meeting enters preliminary review. The dashed lines mark the category-specific population bounds on joint success.

8 Discussion

A randomized comparison of review methods can estimate the difference in their rates of timely recommendations. A statement about which successful review meetings could have been replaced also depends on how the two outcomes align within an item. The alignment is not recovered by a larger sample of one-method-only assignments.

The binary rubric is restrictive. A recommendation recorded after 61 minutes is treated as unsuccessful, and two recommendations that pass the rubric are treated as equivalent. A longer horizon or a quality scale would answer a different question and could change the bounds. The intervention also excludes calls and follow-up meetings during the assessment period. A trial allowing organizers to escalate from email to a meeting would estimate the effect of that protocol, not email alone. These definitions should be settled before randomization.

The independence and no-interference assumptions are strong for shared workplaces. For example, a participant who learns an answer in one meeting may use it in a later email. Calendar selection can exclude cancelled review meetings from a retrospective sample, so the target population should be fixed before assignments are made. The synthetic results establish no prevalence for actual organizations. A study with prospectively defined review items, comparable recommendation rubrics and a design accommodating shared teams would be needed to estimate even the marginal rates.

The recursive model treats review obligations as independent and does not describe how a particular organization schedules or cancels its meetings. Its oracle calculation is a benchmark: an actual organizer cannot observe the two potential outcomes needed to apply that rule. The model nonetheless separates the burden of deciding on a review method from the burden of the meeting under review. The first must be included when considering a policy that applies to review meetings themselves.

A Confidence-envelope optimization

Let the simultaneous rectangle be $m_x \in [\ell_{Mx}, u_{Mx}]$ and $e_x \in [\ell_{Ex}, u_{Ex}]$. Both endpoints of Eq. (3) are nondecreasing in each e_x, so the minimum uses $e_x = \ell_{Ex}$ and the maximum uses $e_x = u_{Ex}$. For the lower endpoint, holding other coordinates fixed, increasing m_x decreases the ratio while $m_x + e_x < 1$ and increases it after $m_x + e_x > 1$. Thus a minimizing m_x is $1 - \ell_{Ex}$ clipped to $[\ell_{Mx}, u_{Mx}]$. For the upper endpoint, the ratio increases while $m_x < e_x$ and decreases while $m_x > e_x$, so a maximizing m_x is u_{Ex} clipped to the same interval. These coordinatewise choices apply simultaneously. If a denominator vanishes, the implementation uses the conservative endpoint 0 or 1. Limits are computed before rounding.

B Data and code availability

The authors retain the simulated observations and Python implementation. No data on real meeting reviews or employees were collected. The following Python 3.14 code regenerates the counts in Table 2. For the 2,000 replication trials, call `trial(rng)` a further 2,000 times using the same generator and apply Eq. (3) and Eq. (4) to each trial. Items are drawn in the order shown in Table 1. The first half after shuffling is assigned to meetings and the rest to email.

```
import random
rng = random.Random(534976)
settings = [(500, .90, .95, .875),
            (250, .96, .72, .700),
            (250, .44, .18, .100)]
def trial(rng):
    result = []
    for n, m, e, q in settings:
        pairs = []
        for _ in range(2*n):
            u = rng.random()
            pairs.append((1, 1) if u < q else
                         (1, 0) if u < m else
                         (0, 1) if u < m+e-q else (0, 0))
        rng.shuffle(pairs)
        result.append((sum(a for a,b in pairs[:n]),
                       sum(b for a,b in pairs[n:])))
    return result
print(trial(rng))
```

References

Wassily Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Guido W. Imbens and Charles F. Manski. Confidence Intervals for Partially Identified Parameters. *Econometrica*, 72(6):1845–1857, 2004. doi: 10.1111/j.1468-0262.2004.00555.x.

Alexandra Luong and Steven G. Rogelberg. Meetings and More Meetings: The Relationship Between Meeting Load and the Daily Well-Being of Employees. *Group Dynamics: Theory, Research, and Practice*, 9(1):58–67, 2005. doi: 10.1037/1089-2699.9.1.58.

Donald B. Rubin. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350.

Jin Tian and Judea Pearl. Probabilities of Causation: Bounds and Identification. *Annals of Mathematics and Artificial Intelligence*, 28:287–313, 2000. doi: 10.1023/A:1018912507879.