Confirming Recall Adequacy With Unbiased Multi-Stage Acceptance Testing

William C. Dimm
Hot Neuron LLC
bdimm@hotneuron.com

ABSTRACT

The adequacy of an e-discovery production has traditionally been established by using random sampling to estimate recall, but that requires review of approximately $400/\rho$ documents, where ρ is the prevalence, which can be burdensome when prevalence is low. Accept-on-zero testing is sometimes suggested as an option requiring less review at only about $12/\rho$, but in practice it is biased and is likely to fail when recall actually is adequate. This paper proposes a multi-stage acceptance testing procedure that avoids bias and actually works in practice. The amount of document review required with the new method depends on the level of recall actually achieved. It is typically around $200/\rho$ or $100/\rho$, but can be as low as $25/\rho$ if the actual recall is substantially higher than the minimum required. This dependence on the recall achieved may motivate producing parties to aim for higher recall since the additional document review put into pushing recall higher will be at least partially offset by a reduction in review effort needed to confirm the adequacy of the result.

1 RECALL ESTIMATION

One can estimate the proportion of all documents having some property by measuring the proportion of a random sample having the property of interest. Recall is the proportion of responsive documents that are actually produced. The approach known as the direct method for estimating recall typically involves selecting approximately 400 random responsive documents from the full population and identifying the proportion that were produced.[1] For example, if 300 of the responsive documents have been reviewed and produced, either during training of a predictive coding system or because the system predicted that they were responsive, whereas 100 of the responsive documents have been incorrectly predicted to be non-responsive by the system and were thus not reviewed or produced, the resulting confidence interval for the recall would be 75% ± 5% with 95% confidence, so we are reasonably sure that the actual recall is between 70% and 80% (this is an approximation; the exact confidence interval computed with the Clopper-Pearson method is 70.5% to 79.2% [CP34]). To be more precise, regardless of what the actual recall is, taking a sample of 400 documents and computing the interval ±5% around the point estimate from the sample will give an interval that captures the actual recall at least 95% of the time. The width of the confidence interval is inversely proportional to the square root of the sample size, so a confidence interval of ±2.5% with 95% confidence would require 1,600 random responsive documents—estimating recall more precisely requires a significantly larger sample.

Finding even 400 random responsive documents to perform the estimate described above can be costly. One cannot sample directly from the set of responsive documents because it is not known which documents in the population are responsive—only a subset that have been reviewed are known. To obtain 400 random responsive documents, one must choose documents randomly from the entire population, review them, and discard the non-responsive ones until 400 responsive documents are found. If prevalence is 10%, that means reviewing approximately 4,000 documents in order to find 400 that are responsive. If prevalence is 1%, review of approximately 40,000 documents is required. In general, approximately $400/\rho$ documents must be reviewed, where ρ is the prevalence. Demonstrating that a production is adequate by estimating recall in this way can involve an amount of document review that is substantial compared to the review that was performed to actually find and produce the responsive documents.

In the past, there have been some suggestions that different approaches could be used to estimate recall with far less document review. Those approaches involved estimating the numerator and denominator of the recall separately using a sample of only 385 documents (not 385 *responsive* documents) each and computing the ratio. Proponents of such approaches apparently didn't understand how uncertainty in a variable in an equation can be amplified into a much larger uncertainty in the final result. Grossman and Cormack showed via simulation that ratio methods utilizing only 770 random documents when prevalence was 1% resulted in estimates with such large uncertainty that they were virtually meaningless [GC14, p. 305]. Instead of estimates being mostly within ±5% of the right result, they were actually off by ±25% or more. Dimm showed mathematically that achieving a correctly-computed confidence interval of ±5% with one of the ratio approaches required a sample size no smaller than the direct method [Dim14]. Furthermore, combining measurements from sampling different parts of the population introduces the possibility of bias if the documents in the samples are not mixed together before they are reviewed because the reviewer may tend to make different mistakes or judgment calls on borderline documents if the samples have very different prevalence.

A recall estimate is not an adequacy determination, and that distinction will be important for the remainder of this paper. A decision must be made about whether the estimated recall is high enough, and proportionality will determine what is considered adequate—a high-value case where e-discovery is critical may demand higher recall than a small case.

[1] People often use a sample size of 385, but that number comes from an approximation applied to the worst case scenario (50% recall), so good arguments could be made for other values. We use 400 throughout to make the numbers easier for the reader to follow.

DESI VII, London, UK
2017. 978-x-xxxx-xxxx-x/YY/MM.
DOI:

Suppose, for the sake of discussion, that finding 300 out of a sample of 400 random responsive documents were produced is the absolute minimum that is acceptable for a particular case that is under consideration. If an attempt was made to cut costs by sampling only 100 random responsive documents instead of 400, and it was found that 75 of the 100 were produced, would that be considered acceptable? Probably not. The point estimate for the recall is the same, 75%, but the confidence interval would be 75% ± 10%. The actual recall might reasonably be expected to be as low as 65%, rather than the 70% lower bound from the 400 document sample. On the other hand, if 80 of the 100 documents were produced the confidence interval would be 80% ± 10% (actually 70.8% to 87.3% when computed exactly), giving a lower bound that is no worse than finding that 300 out of 400 responsive documents were produced. It is reasonable to claim that 80 out of 100 should be considered acceptable if 300 out of 400 is.

Figure 1 compares the probability of accepting the production as adequate for the two criteria considered in the previous paragraph as a function of actual recall achieved.[2] As suggested earlier, the graph shows that the probability of accepting the production as sufficient is very low (less than 2.5%) for both criteria if the actual recall is less than 70%. Where the two criteria differ is in the possibility of failing to accept a production when the recall actually is sufficiently high. For example, if the actual recall is 80%, a sample of 400 random responsive documents would almost certainly (more than 97.5% probability) show that at least 300 of them were produced. Using a sample of only 100 responsive documents, there is only about a 50% chance that at least 80 would be found to have been produced.

Using a smaller sample reduces the amount of review, but increases the risk of incorrectly rejecting the production as inadequate when the actual recall is sufficient. One might contemplate sampling 100 responsive documents, checking to see if at least 80 were produced, and continuing to review documents until a total of 400 had been sampled if the test at 100 documents failed. Such multiple testing increases the risk of accepting a result when the actual recall isn't high enough, but the basic idea is the driver behind the method that will be explained in this paper. The important point is to engineer a multi-stage acceptance test so the probabilities work out, which is more complicated than gluing together two tests that are independently valid but not valid in combination.

2 ACCEPT-ON-ZERO TESTING

Accept-on-zero testing is sometimes mentioned as a way to test adequacy with a much smaller sample [Roi]. If recall is adequate, there should be relatively few responsive documents left among the documents that have not been reviewed or produced (sometimes referred to as the discard set or negatives). The proportion of the discard set that is responsive, known as the elusion or false omission rate, can be confirmed to be below a certain level by reviewing a random sample of a certain size and confirming that none of the documents are responsive. To determine how small the elusion needs to be, we examine the relationship between elusion and recall:

$$E = \rho \frac{1 - R}{1 - \frac{\rho}{P} R} \tag{1}$$

The denominator involves two quantities that are individually unknown, the prevalence and precision, but it is equal to the proportion of the full population that is in the discard set, which is known. That still leaves an overall factor of the prevalence in the equation, which is typically either completely unknown or known relatively imprecisely if some sampling is done and the prevalence happens to be low. We'll ignore that problem and examine the size of the sample that is required to establish adequate recall. We're most interested in small prevalence since that is when estimating recall with the direct method is burdensome, so we'll take the denominator in Equation (1) to be approximately one. The maximum elusion we can tolerate if we want to ensure the recall is above some minimum value is:

$$E_{\max} \approx \rho(1 - R_{\min}) \tag{2}$$

If a production is accepted if a sample of n documents from the discard set turns up none that are responsive, the probability of acceptance is:

$$p_{\text{accept}} = (1 - E)^n \tag{3}$$

To ensure the probability of accepting a production having recall less than $R_{\min}$ is no more than 2.5%, we choose n such that:

$$(1 - E_{\max})^n = 0.025 \tag{4}$$

Giving:

$$n = \frac{\log(0.025)}{\log[1 - \rho(1 - R_{\min})]} \approx \frac{-\log(0.025)}{\rho(1 - R_{\min})} \tag{5}$$

If $R_{\min}$ is 70% (to match the 300 out of 400 responsive documents being produced result, which gives a probability of at most 2.5% of accepting a result where the actual recall is 70%), we have $n \approx 12/\rho$, which is vastly better than the $400/\rho$ sample size used for recall estimation.

Substituting Equation (5) into Equation (3) gives:

$$p_{\text{accept}} = [1 - \rho(1 - R)]^{\frac{-\log(0.025)}{\rho(1 - R_{\min})}} \tag{6}$$

Although this equation depends on prevalence, the dependence is extremely weak. Figure 1 shows that the probability of accepting a result having low recall is small, as expected, but the probability of accepting a production where the recall is high is also small. Unless the actual recall is over 94%, it is more likely that the accept-on-zero test will reject rather than accept a result having high recall. After applying the accept-on-zero test and having it reject the production because one or more of the documents in the sample from the discard set are responsive, what is the next step? As mentioned in the previous section, you cannot apply a test over and over until you get a favorable result or the probability of accepting a production having low recall will grow with each additional application of the test. The accept-on-zero test gives a high probability of being stuck with a negative result even when the production is actually adequate.

The accept-on-zero test involves sampling from the discard set, but that's not much different from taking a slightly larger sample from the full population and expecting 100% of the responsive documents found to be documents that were produced (not in the

[2] Computed by summing the binomial distribution function.

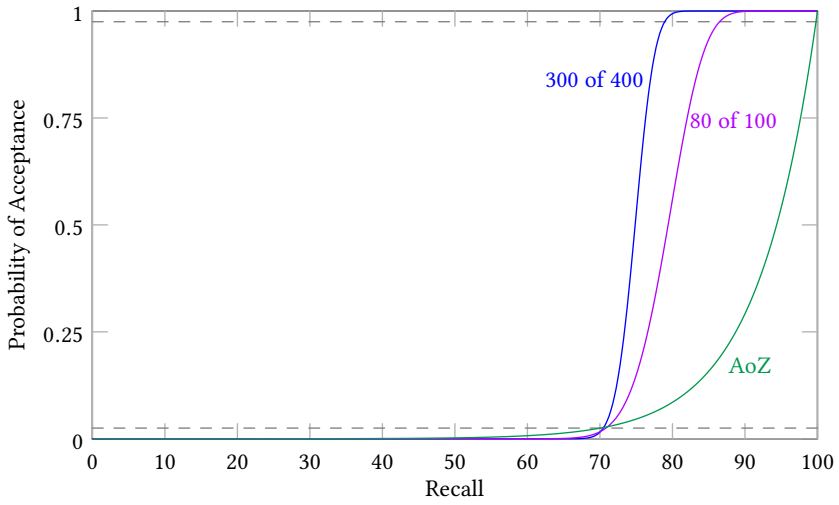

Figure 1: How the probability of acceptance depends on actual recall for three criteria: A sample of 400 responsive documents requiring at least 300 to have been produced, 100 documents with at least 80 produced, and the accept-on-zero test. Dashed lines are at 2.5% and 97.5% probability.

discard set). The previous section and Figure 1 showed the impact of shrinking the sample from 400 documents to 100 and requiring that a higher percentage of the responsive documents found were documents that were produced (80% instead of 75%). The accept-on-zero test just takes that idea to the extreme, which results in such a large risk of rejecting a production where the recall actually is high that it is useless. The test can be modified to use a larger sample and allow acceptance with one or even several responsive documents found in the sample, but that just brings the test back toward where we started with the large samples used for recall estimation [L$^+$15, sect. 9].

Bias is also a problem. Declaring a single document in the sample having borderline responsiveness to be non-responsive can flip the entire outcome of the test from rejection to acceptance, so the test is very sensitive to any shift in the reviewer's concept of responsiveness or any tendency to mistakenly mark responsive documents as non-responsive when the vast majority of the documents being reviewed are non-responsive.[3] Sampling from the full population, rather than from the discard set, ensures that any shift in the concept of responsiveness impacts all documents, whereas sampling from the discard set makes the impact completely one-sided.

3 MULTI-STAGE ACCEPTANCE TESTING

This approach samples from the full population rather than the discard set in order to avoid the bias problem discussed in the previous section. A splitting recall, which we'll denote R_s, is chosen with the aim of accepting productions where the actual recall is above R_s and rejecting productions where the actual recall is below R_s. Of course, decisions will be made based on sampling, so we cannot expect perfection if the actual recall is very close to R_s. The first stage involves a relatively small sample and accepts or rejects the production if the proportion of the sample that was produced is far above or far below R_s, respectively. If the proportion of the sample that was produced is too close to R_s to make a decision without significant risk of being wrong about where the actual recall lies relative to R_s, the sample is enlarged and tested against a tighter set of boundaries with the possibility of iterating several times. The accept/reject boundaries are engineered so the total probability of making a bad adequacy determination for the entire process is controlled (there is no multiple testing problem), and the sample sizes are optimized to minimize the amount of document review.

Figure 2 is a flowchart illustrating a multi-stage acceptance testing procedure for $R_s = 75\%$, which is intended to give results similar to requiring a sample of 400 responsive documents to have at least 300 that were produced, but with much less document review. Figure 3 shows that the probability of acceptance as a function of actual recall achieved is extremely similar to the result from requiring at least 300 out of 400 responsive documents to have been produced. Figure 4 shows that the amount of document review required is much less than the 400 responsive documents used for recall estimation. The amount of review required depends on the actual recall that was achieved. When the actual recall is far above/below R_s, the amount of review is especially small because the production will typically be be accepted/rejected during the early stages of the procedure.

The multi-stage procedure is designed so that the risk of drawing the wrong conclusion about the adequacy of the production is very small when the actual recall is below $R_s - 5\%$ (very likely rejected) or above $R_s + 5\%$ (very likely accepted). When the actual recall is equal to R_s, the probability of accepting the production is approximately 50%. The amount of document review required is largest close to R_s. These are all factors that should be kept in mind when choosing an R_s that is appropriate for a case. If 75% recall is acceptable, you may want to set R_s to 70%, or even lower if the requesting

[3] This was pointed about by David Lewis in a private communication, which inspired this research.

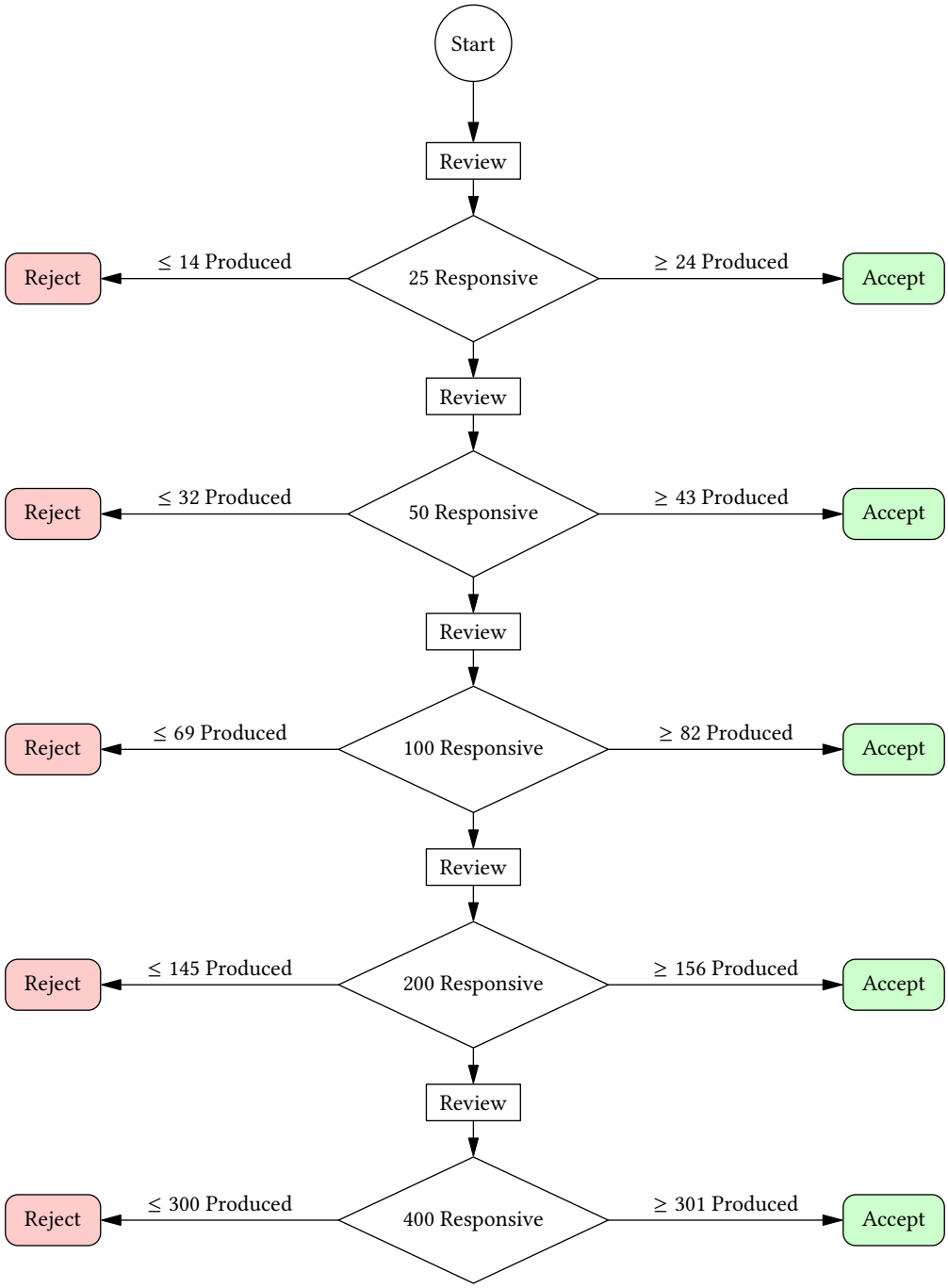

Figure 2: Multi-stage acceptance test with splitting recall of 75%. The probability of incorrectly accepting the production if the actual recall is less than 70% is at most 2.5%. The probability of incorrectly rejecting the production if the actual recall is greater than 80% is at most 2.5%.

party is amenable, to ensure that a production with 75% recall is virtually always accepted (instead of being accepted half of the time, as it would be if R_s was 75%) and to benefit from a greater reduction in review effort required to test the production when actual recall is well above R_s. Table 1 specifies the parameter values for applying the multi-stage method with many different R_s values. Table 2 shows the average amount of document review required for various R_s values and levels of actual recall.

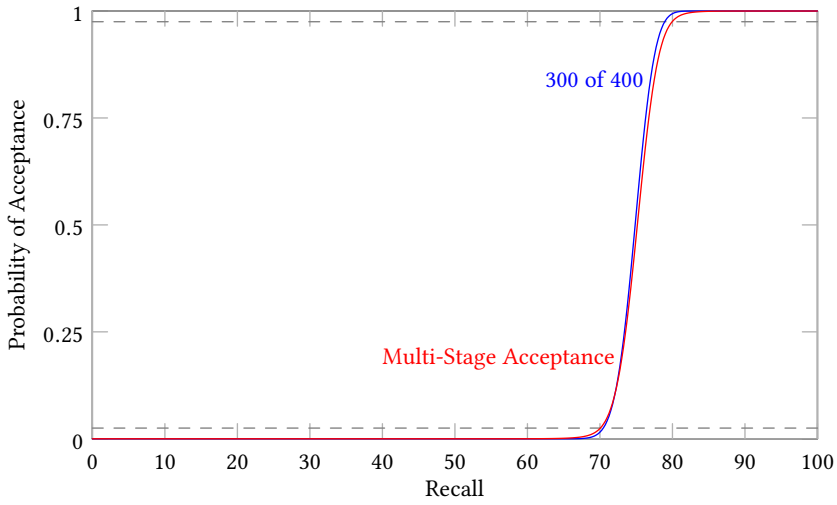

Figure 3: Comparison of multi-stage acceptance testing procedure from Figure 2 ($R_s = 75\%$) to requiring at least 300 documents from a 400 responsive document sample to have been produced.

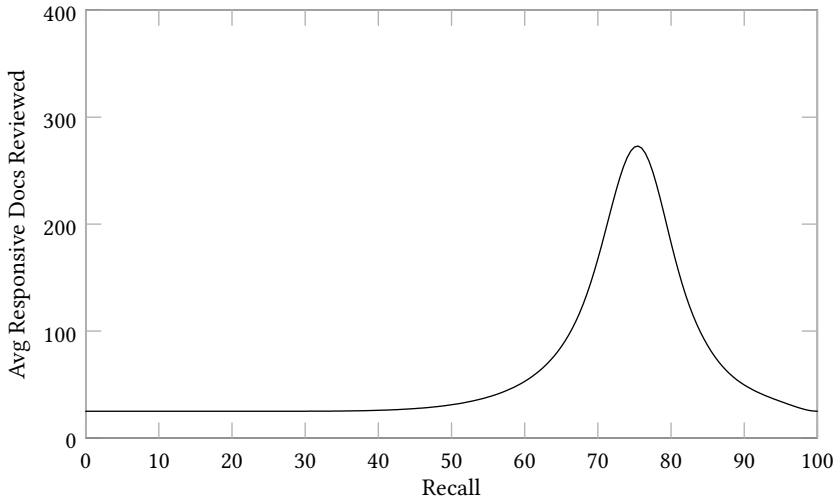

Figure 4: Multi-stage acceptance testing procedure from Figure 2 ($R_s = 75\%$) requires much less document review than the 400 responsive documents that were needed to estimate recall.

	R_s													
Samples	60%		65%		70%		75%		80%		85%		90%	
25	8	21	9	22	11	23	14	24	16	25	17	25	20	25
50	22	38	26	40	29	41	32	43	35	45	39	47	43	49
100	50	70	56	74	63	78	69	82	75	85	82	90	88	94
200	111	129	122	138	134	148	145	156	157	165	169	173	181	183
400	240	241	260	261	280	281	300	301	320	321	340	341	360	361
Rev(R_s)	339.8		322.6		298.9		272.1		234.8		185.5		136.1	
Avg Rev	137.0		127.7		116.1		105.5		94.1		81.1		62.3	

Table 1: Rejection and acceptance boundaries (inclusive) for the number of sample responsive documents shown on the left and various values of the splitting recall, R_s, indicated along the top. Probability of drawing an incorrect conclusion when the actual recall is outside $R_s \pm 5\%$ is at most 2.5%. The average number of responsive documents that must be reviewed when the actual recall is R_s is shown along the bottom—this indicates the review effort required when the actual recall is least optimal. Below that, the weighted average review across a spectrum of actual recall levels is shown.

	R_s						
Actual Recall	60%	65%	70%	75%	80%	85%	90%
0%	25.0	25.0	25.0	25.0	25.0	25.0	25.0
5%	25.0	25.0	25.0	25.0	25.0	25.0	25.0
10%	25.0	25.0	25.0	25.0	25.0	25.0	25.0
15%	25.2	25.1	25.0	25.0	25.0	25.0	25.0
20%	26.2	25.4	25.0	25.0	25.0	25.0	25.0
25%	28.8	26.8	25.3	25.0	25.0	25.0	25.0
30%	33.7	29.8	26.1	25.0	25.0	25.0	25.0
35%	41.8	34.5	28.1	25.2	25.0	25.0	25.0
40%	56.1	41.0	31.9	25.9	25.1	25.0	25.0
45%	83.8	51.2	37.6	27.5	25.4	25.1	25.0
50%	139.3	72.3	46.7	31.1	26.4	25.5	25.0
55%	252.3	120.3	63.5	38.3	28.8	26.6	25.1
60%	339.8	227.8	100.6	52.9	34.3	28.9	25.2
65%	251.8	322.6	191.8	85.4	46.2	33.4	25.8
70%	135.2	232.4	298.9	167.5	72.9	41.8	27.4
75%	77.6	121.2	214.5	272.1	138.4	59.3	31.2
80%	48.8	69.2	100.9	182.6	234.8	104.2	41.1
85%	34.3	44.0	55.8	86.5	148.1	185.5	70.4
90%	27.5	31.3	37.8	49.6	71.0	126.6	136.1
95%	25.2	25.9	28.2	34.1	44.9	55.6	93.0
100%	25.0	25.0	25.0	25.0	25.0	25.0	25.0

Table 2: Average number of responsive documents that must be reviewed for various levels of actual recall listed on the left for various values of the splitting recall, R_s, indicated along the top. Probability of drawing an incorrect conclusion when the actual recall is outside $R_s \pm 5\%$ is at most 2.5%.

	R_s													
Samples	60%		65%		70%		75%		80%		85%		90%	
24	8	20	10	21	12	22	13	22	15	23	17	24	20	24
45	19	34	23	35	26	37	29	39	32	40	35	42	39	44
83	42	58	47	61	52	64	58	68	63	71	68	74	73	78
153	84	99	93	107	102	113	111	120	120	127	130	132	138	0
280	168	169	182	183	196	197	210	211	224	225	238	239	252	253
Rev(R_s)	232.3		218.1		198.3		179.6		156.3		115.1		85.3	
Avg Rev	104.0		94.5		86.2		79.1		70.5		59.8		46.2	

Table 3: Rejection and acceptance boundaries (inclusive) for the number of sample responsive documents shown on the left and various values of the splitting recall, R_s, indicated along the top. Probability of drawing an incorrect conclusion when the actual recall is outside $R_s \pm 5\%$ is at most 5%. The average number of responsive documents that must be reviewed when the actual recall is R_s is shown along the bottom—this indicates the review effort required when the actual recall is least optimal. Below that, the weighted average review across a spectrum of actual recall levels is shown.

	R_s						
Actual Recall	60%	65%	70%	75%	80%	85%	90%
0%	24.0	24.0	24.0	24.0	24.0	24.0	24.0
5%	24.0	24.0	24.0	24.0	24.0	24.0	24.0
10%	24.0	24.0	24.0	24.0	24.0	24.0	24.0
15%	24.1	24.0	24.0	24.0	24.0	24.0	24.0
20%	24.8	24.1	24.0	24.0	24.0	24.0	24.0
25%	26.7	24.4	24.0	24.0	24.0	24.0	24.0
30%	30.8	25.6	24.2	24.1	24.0	24.0	24.0
35%	38.4	28.2	24.9	24.3	24.0	24.0	24.0
40%	51.0	33.1	26.6	25.1	24.2	24.0	24.0
45%	72.1	42.2	30.2	26.9	24.6	24.1	24.0
50%	112.9	59.5	37.4	30.3	25.6	24.2	24.0
55%	186.2	95.1	51.7	36.4	28.0	24.8	24.0
60%	232.3	164.0	82.0	47.5	32.5	26.1	24.1
65%	181.4	218.1	143.9	69.8	41.4	29.2	24.3
70%	106.1	170.2	198.3	120.9	59.5	35.8	25.0
75%	63.8	91.7	148.8	179.6	101.2	49.9	27.0
80%	42.2	53.4	79.6	136.1	156.3	79.9	33.1
85%	31.1	36.9	48.7	69.1	113.1	115.1	51.7
90%	25.8	28.6	34.3	39.0	53.8	88.2	85.3
95%	24.1	24.6	26.4	26.7	32.0	46.3	70.5
100%	24.0	24.0	24.0	24.0	24.0	24.0	24.0

Table 4: Average number of responsive documents that must be reviewed for various levels of actual recall listed on the left for various values of the splitting recall, R_s, indicated along the top. Probability of drawing an incorrect conclusion when the actual recall is outside $R_s \pm 5\%$ is at most 5%.

The results presented so far have assumed the risk of making a bad acceptance determination should be less than 2.5% when the actual recall is below $R_s - 5\%$ or above $R_s + 5\%$ for consistency with the standard of performing recall estimation with ±5% confidence intervals and 95% confidence (hence the excellent agreement in Figure 3). If 95% confidence means that one is willing to be wrong 5% of the time, it may be worth considering allowing the acceptance test to be wrong 5% of the time instead of 2.5% when the actual recall is below $R_s - 5\%$ or above $R_s + 5\%$. Table 3 provides the parameter values for allowing a 5% error rate, and Table 4 shows the corresponding amount of document review required.

Parameter values were determined by minimizing the weighted average amount of document review required for various equally-spaced levels of actual recall with constraints on the error rate at $R_s - 5\%$ and $R_s + 5\%$. The weights applied were from a Gaussian distribution centered at R_s with standard deviation of 20%, since actual recall is expected to be somewhat close to the chosen R_s in practice. The weighted average amount of document review is shown at the bottom of Tables 1 and 3.

Computations were performed numerically to evolve probabilities based on the binomial distribution through the stages of the acceptance test. Various experiment were performed where the number of stages was varied, the number of samples at each stage was varied, and the number of samples for accept/reject at the first stage were allowed to vary independently. We ultimately decided that five stages using the same number of samples for all R_s was a sensible trade-off between simplicity and reaching the absolute lowest amount of document review. We also intentionally kept the number of samples in the final stage close to the the number of samples that would be used for traditional recall estimation so that someone who was unlucky enough to reach the final stage would not be penalized compared to doing a simple recall estimate.

4 CONCLUSIONS

The multi-stage acceptance test provides a significant reduction in the amount of document review required to confirm the adequacy of a production. Furthermore, the reduction in effort is larger the farther above the selected splitting recall, R_s, the actual recall is. This benefits e-discovery by encouraging the shifting of effort toward finding more responsive documents instead of putting that effort into confirming the result.

It is worth noting, however, that this procedure gives an accept/reject result, not a recall estimate, and it would not be valid to compute a recall estimate from the documents analyzed during the procedure. If the procedure is terminated because a boundary is hit that causes the production to be accepted, a recall estimate based on that sample would be biased upward because a random upward fluctuation in the proportion produced would increase the chances of hitting that boundary. Likewise, if the procedure is terminated because a rejection boundary is hit, a recall estimate based on the reviewed documents would be biased downward.

REFERENCES

[CP34] C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

[Dim14] William C. Dimm. eRecall: No free lunch. September 2014. https://blog.cluster-text.com/2014/09/08/erecall-no-free-lunch/.

[GC14] Maura R. Grossman and Gordon V. Cormack. Comments on "The implications of rule 26(g) on the use of technology-assisted review". *The Federal Courts Law Review*, 7:285–313, 2014.

[L+15] Michael Levine et al. EDRM statistical sampling applied to electronic discovery. February 2015. http://www.edrm.net/resources/project-guides/edrm-statistical-sampling-applied-to-electronic-discovery/.

[Roi] Herbert L. Roitblat. Measurement in ediscovery: A technical white paper. *OrcaTec*.

Received May 2017