Relativity®

White Paper

Control Sets: Introducing Precision, Recall, and F1 into Relativity Assisted Review

Contents

What Are Control Sets in Assisted Review?

"Okay, students, please take out some paper and a pencil for a pop quiz."

This is perhaps one of the most dreaded phrases uttered by teachers to students. A pop quiz is an unbiased test of a student's knowledge of a subject, and is used as a benchmark by teachers to see how the student is progressing. The key to the pop quiz is that the exact answers to the quiz were not previously provided by the teacher. This way, the teacher is able to see if the student is actually learning versus memorizing.

The new control set feature for Relativity Assisted Review is a lot like a pop quiz is to a teacher—an unbiased and effective way to measure how your system is progressing in its learning throughout the Assisted Review process that makes sure you're the only one with the answer key.

Let's first take a look at how control sets work, and then apply them to an example workflow.

Defining Control Sets

In research, the scientific method is the process used for testing ideas and theories via experiments and careful observation. The procedure for a control group—or **control set**—exists within the scientific method, as part of experimentation. When conducting an experiment, a control group is identified and set aside, unaffected by the experiment's progress, so that it is free of variables and will not change. The control group then serves as a baseline against which all results are evaluated.

Below are the steps of the scientific method as control groups would be incorporated:

- Ask a Question
- Do Background Research
- Construct a hypothesis
- *Set Aside a Control Group*
- Test Your Hypothesis by Doing an Experiment
- *Analyze Your Data by Comparing Results to the Control Group*
- Draw a Conclusion
- Report Results

The control set in Assisted Review mimics this. It is a random sample of documents drawn from the entire collection of documents, usually prior to starting Assisted Review training rounds. Like other sample sets in Assisted Review, the review team's selection of statistical variables and the size of the document universe will determine the size of the control set.

The control set is coded by domain experts for responsiveness and key issues. As mentioned in our white paper validating the computer-assisted review workflow, the coded control set is now considered the human-selected ground truth set and used as a benchmark for further statistical measurements we may want to calculate later in the project. As a result, there is only one active control set in Assisted Review for any given project.

Similar to the teacher never giving the answers prior to the pop quiz, control set documents are never provided to the analytics engine as example documents. Because of this approach, we are able to see how the analytics engine categorizes the control set documents based on its learning, and calculate how well the engine is performing at the end of a particular round. The control set, regardless of size or type, will always be evaluated at the end of every round—a pop quiz for Assisted Review. This gives the Assisted Review team a great deal of flexibility in training the engine, while still using statistics to report on the efficacy of the Assisted Review process.

> *"The control set, regardless of size or type, will always be evaluated at the end of every round—a pop quiz for Assisted Review."*

What Are Precision, Recall, and F1—and How Do We Use Them?

As stated in Measuring and Validating the Effectiveness of Relativity Assisted Review, the field of information retrieval has defined two core metrics for assessing the effectiveness of a search or document categorization workflow—both

of which require both responsive and non-responsive categorized documents to be determined. The first is **precision**—the ratio of all documents categorized as responsive to the number of those documents that were categorized correctly. If a collection has low precision, this means there are a number of non-responsive documents categorized as responsive, demonstrating that the computer's decisions are not very accurate yet.

Recall is the ratio of responsive documents found and categorized correctly, to the total number of responsive documents in the full collection. If a collection has low recall, this means that the computer has not yet found a number of responsive documents.

There is a way to guarantee 100 percent recall: manually review every document with perfect human decision-making. However, as we all understand, this would be inefficient, nearly impossible, and costly. We want to use process and technology like Assisted Review and allow for this trade-off relationship between precision and recall. Luckily, information retrieval has created a metric that balances this trade-off. The **F1** measurement is the harmonic mean—a weighted average of precision and recall—with the metric slightly leaning toward rewarding higher recall.

All in all, the use of precision will approximate how accurate Assisted Review is when it categorizes a document as responsive. The use of recall will approximate the percentage of responsive documents Assisted Review found based on the computer's analytics model.

Why Do I Need to Use a Control Set in Assisted Review to Determine Precision, Recall, and F1?

The Assisted Review workflow was built from the ground-up to have quality control and statistical sampling as core tenets of the process. Appendix A outlines how—to use metrics such as precision, recall, and F1—there must be ground-truth responsive documents included in the measurement. An unbiased, statistically significant control set will allow the Assisted Review team to always have the proper ingredients to display precision, recall, and F1 at any point in the project.

It is important to understand that the control set is providing metrics about the universe from which it was originally sampled. If during the course of the project more documents are imported into a case, the current control set cannot be relied upon for accurate statistics. *In this instance, a new control set should be created.*

How Control Sets Allow for Flexible Training

Many Assisted Review users ask, "When should we stop training the computer and move to QC rounds?" By using a control set workflow, users can continually train and QC, while still keeping their eyes on the overall project efficacy metrics described in the preceding section.

Once the control set is already reviewed and set aside to avoid biasing the analytics engine driving Assisted Review, the project manager has complete flexibility for the size and population of the seed set for training. The Assisted Review manager now has the ability to provide batches of documents based on factors rooted in more than statistical variables, e.g. the number of reviewers or documents per hour.

What about Quality Control and Stratification?

In statistical sampling, there may be times when only a portion of the document universe is desired for analysis and used to extrapolate information. This is called **stratification**, the process of randomly sampling from distinct pools of data via specific criteria. Using a control set and overturn workflow will allow for stratification, while also letting you monitor the progress of the project.

If we were to sample only categorized responsive documents and measure the overturns—the number of times reviewers disagreed with the computer's decisions, overturning them—we would only be able to extrapolate on the overturns for the categorized responsive population. However, if we have a control set, we can estimate an answer on the overturn percentage for the categorized responsive population and use the metrics from the control set to estimate precision and recall on the overall document universe. This scenario may happen when a review team needs to roll out a production quickly and wants to focus on a specific sub-section of the universe on a priority basis.

Sample Walkthrough of a Control Set Workflow

Understanding the use and benefits of control sets can lead to interesting workflows within Relativity. Let's look at a hypothetical Assisted Review project.

In this case[1], we have roughly 500,000 documents comprised of emails and attachments. We have already created an analytics index, and made sure to have the proper filters and searches prepared to optimize the index. Additionally, we have only one domain expert to work on the review.

Starting with a Control Set

We decide to use control sets, and to report on precision, recall, and F1 metrics to the review team.

To start the process, we create a new round with a control set as the specified type. We select a confidence level of 95 percent and margin of error of 5 percent. Assisted Review randomly pulls 384 documents from across our document universe into a control set. We now assign out the documents for the domain expert to review for responsiveness. The domain expert will also assign key issues to any document that is responsive.

Taking a look at the coding decisions for the control set, we see that the domain expert designated 122 documents as responsive (32 percent of the sample). Because this was a random sample across the population, we would expect the overall universe of documents to also be around 32 percent responsive.

Table 1: Control Set Coding Decisions

Reviewed Docs in Control Set	Confidence Level	Margin of Error	Coded Responsive	Coded Non-Responsive
384	95%	5%	122 (31.77%)	262 (68.22%)

1 The case in this section is fictional and is for demonstration purposes only. However, it is based on experience with these case types, and is intended to be representative of a case that would likely benefit from this workflow.

Beginning the First Training Phase

We explain to the domain expert that we will use the control set as a baseline to see how the analytics engine progresses from round to round, and we will not use these documents to train the analytics engine. The expert agrees, but wants to start training and see how the computer is progressing before leaving for the day.

Our expert finished reviewing the control set's 384 documents in approximately five hours, averaging 75 documents per hour. With this information, we create the first training round with a fixed amount of 225 documents. This will allow enough time for the reviewer to finish reviewing a batch of documents and also for us to report on the metrics at the end of the day. As in all rounds, the documents are randomly pulled from the overall document universe based on the selected saved search.

After the reviewer finishes the round, Assisted Review categorizes the documents and creates reports automatically. After the first training round based on 225 documents, the control set statistics and the round summary report read:

Table 2: Control Set Statistics after First Training Round

Precision	Recall	F1	Categorized Responsive	Categorized Non-Responsive	Uncategorized
33.78%	43.10%	37.88%	24.54%	64.31%	11.15%

Table 3: Round Summary Report

Categorized Responsive	Categorized Non-Responsive	Uncategorized
89,360 (22.34%)	263,880 (65.97%)	46,760 (11.69%)

The report is telling us that in both the control set and the overall project, around 11 percent of documents are uncategorized. We can also interpret from the control set report that the engine is finding only a little more than four out of every 10 responsive documents (43.10 percent recall). The report is also saying that, of the documents categorized as responsive, only three out of every 10 are

correctly identified (33.78 percent precision). The analytics engine obviously needs additional training.

We will continue to create training rounds until the uncategorized percentage is at a manageable level for a manual review work stream.

Moving on to Training Phase II

With control sets, the focus is on making upward progress on precision, recall, and F1. We will use the overturn report sparingly, primarily to confirm that the control set statistics apply to the overall document universe.

At this point, we will create training rounds of 75 documents, roughly an hour of document review for our domain expert. This will allow us to review and categorize quickly, and be able to focus on specific areas where we want to improve the metrics. For example, if we need to improve precision, we may focus on creating rounds of responsive documents with a high rank score. If we need to focus on improving recall, we may create rounds focused on documents categorized as not responsive that perhaps have high key issue scores. We will continue creating rounds until our control set metrics match the expectations of the review team.

At the end of post-categorized training, our control set metrics read:

Table 4: Control Set Statistics at the End of Post-categorized Training

Precision	Recall	F1	Categorized Responsive	Categorized Non-Responsive	Uncate-gorized
88.78%	91.10%	89.93%	30.78%	66.07%	3.15%

Finishing with the Overturn Report

After reviewing our volatility report (see Figure 1) and determining that training has most likely plateaued, we confer with the review team. We decide that the precision score is acceptable, and decide we may batch out the categorized responsive documents to our production review team, where they will review for production and privilege.

We decide to conduct a QC round on non-responsive documents to have an official overturn report before closing out the Assisted Review project.

For this QC round, we select non-responsive documents, and use a confidence level of 95 percent and a margin of error of 2.5 percent. Assisted Review randomly selects 1,530 documents that were categorized as not responsive. It should take a little less than three days for the reviewer, at the current review rate, to complete review of documents in this round.

At the end of the round, we review the overturn report with the review team. The overturn summary reports that the overturn range for this round was between 1.5 and 6.5 percent. Because the overturn report tracked alongside the control set statistics, the review team felt comfortable concluding that Assisted Review was complete.

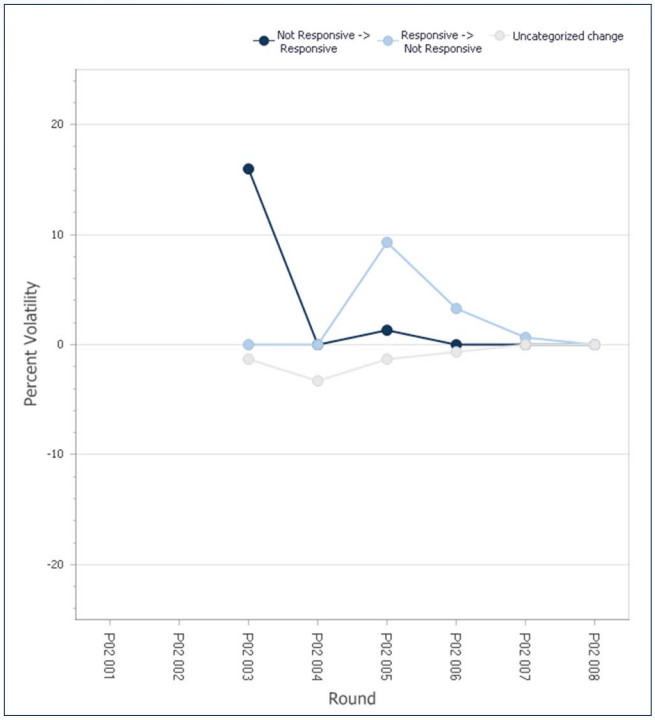

Figure 1: Volatility Report

Final Thoughts

As you can see from this sample workflow, using a control set provides a great deal of flexibility. The pop quiz nature of it allows for Assisted Review administrators to provide metrics for rounds of varying size and settings.

To help guide you through the workflow in more general terms, here's a checklist of the steps for using control sets in your Assisted Review projects:

1. **Create a control group using the control set round type in Assisted Review.**
   Be sure to sample from the full population to be categorized by Assisted Review.

2. **Have a domain expert review the sample for responsiveness and issues.**
   Do not use these documents as examples.

3. **Take note of your control set's percentage of responsive documents.**
   This number will serve as a benchmark throughout your project.

4. **Perform a training round.**
   Because your control set provides benchmarks for comparison, choose the sample size and target population that fits your workflow.

5. **Check your volatility, precision, recall, and F1 metrics to track your progress.**
   You should expect to see an increase in precision, recall, and F1 from round to round.

   Volatility should decrease from round to round.

6. **Repeat steps 4 and 5 until you're satisfied with the metrics you've achieved.**

7. **Your Assisted Review project is complete.**

This workflow can be adapted to meet a variety of case objectives, giving case teams the flexibility to adjust their approach and reference statistically sound metrics that allow for continuous evaluation.

Review teams using Assisted Review have a variety of review goals to meet. They can use the control set metrics as the benchmark for overall project progress, while still accomplishing the unique goals of the project.

kCura®

231 South LaSalle Street, 8th Floor, Chicago, IL 60604
T: 312.263.1177 • F: 312.263.4351
sales@kcura.com • www.kcura.com

Appendix A – Control Set Infographic