Qualitative Disaggregation — A Job Aid

In plain language: we sort what people told us by who they are, then read each group's words.

We are looking for whether the service works differently for different people.

Here's an aid to help you compare how different groups experience your service, in their own words.

Open it when you have a pile of comments and a research question you'd like to answer.

It carries the check we ran in the workshop, the four questions that go with it, and the method around them. In our course, we asked "Who does this service fail?"

Have We Heard Enough? The Hold-Back Test

Hold about one in five of your customer experience data back from your initial read, and never fewer than five items. When you are done, read the hold-back items. Do they add a reason you have not yet encountered?

> In the workshop: hold some back, read them last, ask whether they add a reason you didn't already have.

1. **Set aside a random portion of the comments: about one in five, and never fewer than five. Do not read them.**
2. **Analyze the rest.** Write down every distinct reason the service worked or failed for someone.
3. **Now read the ones you set aside.**
4. **Ask one question: does this contain a reason we did not already have?**

Nothing new? We have evidence, not a feeling, that we have heard the range of what this data holds. For this group, on this question.

Something new? We are not done. And now we know it, instead of assuming it.

That is the difference between a claim and a test. A claim we assert. A test we can fail. That is the only reason passing it means anything.

Two honest limits.

Small piles. If we only have a dozen comments, the test is weak no matter how we split them. Run it anyway, and say so rather than dress it up.

New reasons only. This test is good at catching a reason we had not heard. It is blunter about deeper understanding of a reason we already had. That is real learning too, and harder to check.

Researchers call the state this test looks for *saturation*: the point where more data stops teaching us anything new. The set we hold back is a *holdout*.

The Four Saturation Questions

People will offer a number of interviews that counts as enough. Someone else's number is not evidence about our data. The right number depends on how alike our people are and how narrow our question is. So do not ask "how many is enough?" It has no answer.

Ask these four questions instead. We can answer all of them with the data we already have.

Question 1 — Saturated on What?

We never saturate in general. We saturate on **one question**, about **one group**.

"Have we heard the ways this process fails people who work for themselves?" Answerable.
"Have we heard how people who work for themselves experience government?" Not answerable, ever.

Narrow question, similar group: we will get there quickly. Broad question, mixed group: we may never get there, and we should say so.

Write down the question before you start. If it will not fit in one sentence, it is not yet a question we can saturate on.

Question 2 — When Did the Last New Thing Appear?

We can only answer this if we **read in batches and wrote down what each batch added.**

Read six comments. Write down every distinct reason the service worked or failed for someone. Read six more. Write down anything *new*. Keep going.

If we read the whole pile in one go, we cannot answer this question. Not because we did bad work. Saturation is about the order we read in, not the size of the pile. There is no way to recover it afterward.

This is the cheapest discipline in the whole method, and almost nobody does it.

Question 3 — If We Read More, Would We Learn Something New? Test It.

Run the Hold-Back Test above. Do not guess at the answer. A guess is a claim. The test can fail, which is why passing it counts.

Question 4 — What Could We Never Have Learned from This Data?

We can be perfectly saturated and perfectly blind.

Saturation tells us what **our data** has run out of. It tells us nothing about **who never got into our data.** If our survey only reached people with an email address, we can hear the same thing from every one of them and still miss an entire experience. That experience belongs only to the people without one.

Saturation is never evidence of coverage. Say so, every time.

Count Reasons, Not Themes

Themes multiply with how finely we code. Reasons do not.

When we ask the saturation question, the thing we need the full range of is **the distinct reasons the service works or fails for someone**. Each distinct reason needs a different response.

The test for whether two comments are the same reason: *would the same change address both?* Same change, same reason. Different change, new reason.

Count the ones that work, not only the ones that fail. A person who got through easily is telling us something new if they got through for a reason we had not yet heard. That is often the most useful item in the pile. It tells us what the real barrier was.

Where this sits. Affinity mapping produces *themes*, and themes are the right unit there. The saturation question in disaggregation is a different question: *have we heard the reasons this works or fails for this group?* Different question, different unit. Both are correct in their place.

The Statement to Write

> "We read [the comments] in batches and stopped finding new reasons after [point]. We held back [portion] and read them last. They contained no reason we had not already found. That gives us reasonable confidence we have described the range of reasons this process works and fails for [group].
>
> It gives us no basis for saying how many people are affected, and no information about the people who never reached us."

That statement is our permission and our limit in one breath. It is also a statement we can defend to a skeptical director, because everything in it is something we did.

Our statement, for our data:

Write-on lines. Print this page, or write your statement in your own notes.

Saturation Is Not Prevalence

These get mixed up because both arrive as numbers. They are not the same kind of claim. *Prevalence* means how common something is across a whole population.

Saturation and Prevalence, Compared

Question	Saturation	Prevalence
Answers	Have we heard the range of things there are to hear?	How many people does this affect?
Is a claim about	The set of experiences	The population
We get there by	Listening until people stop telling us new things	Having a sample we can defend
More time with the same comments	Helps	Does not help at all

Think of it this way: saturation is a vocabulary list. Prevalence is a census.

If we want to know what words are spoken in a town, we can stop asking once people stop giving us new ones. If we want to know how many people speak each word, no amount of careful listening will get us there.

What the Method Is

Qualitative disaggregation means comparing how different groups experience a service, using the words people gave us rather than counts.

We do it when we want to know whether a service works differently for different people, and we do not have enough people in each group to compare them with numbers. *Disaggregation* just means breaking a whole into its groups.

Most agencies, boards, commissions, cities, counties, and utilities have small numbers and rich words. This method is built for exactly that.

When to Use It

- Our groups are too small for a defensible statistical comparison, usually under about 30 people per group
- Our question is about **experience**, not about **how many**
- What we have is words: survey comments, complaints, call notes, comment cards, interview notes

When Not to Use It

- **Our question is about frequency.** “How many people had this problem?” is not a question this method answers.

Numbers first, when we can. If we have enough people in each group to compare rates, compare the rates. Numbers give us something words never do: a size.

For example: if we have 200 first-time applicants, 60 of them sole proprietors, we can report that sole proprietors were approved at 50% and everyone else at 80%. That finding has a size, and a director can act on it.

If we have 6 sole proprietors, three of six is 50%, and one more approval makes it 67%. That is a few people in disguise. Read their words instead.

Either way, read the words next. The count finds the gap; reading finds the cause. In our workshop, the count said sole proprietors struggled with the form. Reading said the form was fine. The signature was the problem.

How to Do It

1. **Read everything once, without sorting.** Notice who is talking and what they are talking about.
2. **Work out what our groups actually are.** Not the groups we assumed. The ones the data can support.
3. **Check the *denominator***, the set of people who were even eligible to have this experience. A step that only first-time applicants go through cannot be compared across people who never went through it.
4. **Group the comments by the characteristic we are examining.**
5. **Within each group, look for patterns.**
6. **Across groups, look for differences.** In what people say, and in how much it cost them.
7. **Apply the five rigor moves below.** Every finding gets all five considered, even if we cannot do all five.
8. **Write down what our finding rests on, and what it does not prove.**

Step 6 is where most people go wrong. They count how often each group mentions a thing. That is not disaggregation. Two groups can mention the same problem at the same rate and mean completely different things by it. For one group it is an annoyance. For the other it is the end of the road. **We have to read for meaning, not count for frequency.**

The Five Rigor Moves

Use these as a checklist. If we cannot do one, say so in the write-up. Saying so is part of the method.

1. *Triangulation*. Look for the same finding in more than one place: a different data source, a different method, a different group of people. A pattern that shows up in our survey *and* in our call notes is stronger than one that shows up in either alone.

2. Saturation reasoning. Ask whether new material is still teaching us new things, and then **test it** rather than assume it. There is no number that tells us when we are done. There is a test, and we can run it. See the Hold-Back Test above.

3. *Negative case analysis*. Go looking for the people who *do not* fit our finding. This does not weaken the finding. It usually sharpens it, because the exceptions tell us what the real cause is.

4. Methodological transparency. Every finding gets a plain sentence saying where it came from and what it does not prove. Write it before someone asks for it.

5. *Member checking*. Where we can, take the finding back to the people it describes, or to an organization that works with them, and ask whether it rings true. Where we cannot, say that we could not.

Duplicates, and Why They Inflate a Theme

Take out the duplicates before looking for patterns. Here is why it matters more than it sounds.

A theme is built on independent voices agreeing. Two people saying the same thing is corroboration. One person saying it twice is one person saying it twice. A duplicate creates the appearance of agreement where there is none, and that appearance is the only thing making it a theme.

Duplicates are not random. They cluster. People submit twice when they are angry, when the connection drops, when the page does not respond, and above all **when they cannot tell whether the first one went through.**

Sit with that. If our service does not confirm that it received something, the people most likely to send it twice are **the people our service already failed.** So duplicates do not inflate our themes evenly. They inflate the themes belonging to the people who were treated worst. That is exactly the group we were trying to see clearly.

That is a bias, not noise.

The ones we cannot catch are the dangerous ones. An exact duplicate is easy to spot. But one person who calls, then emails, then fills in the survey produces three items that look independent, sound different, and sit in three different places. They are **one person.** Nothing in our data will tell us that.

We can remove the duplicates we can see. We cannot remove the ones we cannot. **Say so in the write-up.** That is not a weakness in the work. It is the work being honest, which is the only thing that makes it usable.

The reason to bother even when it would not change the answer:

> Every count we produce rests on a decision about what counts as one thing.

If someone finds a duplicate in our data after we have presented our findings, it will not matter that the conclusion was right. What we lose is not that finding. It is our standing to bring the next one.

Three Ways This Goes Wrong

1. Confirmation bias dressed up as a theme.
We already believe something about a group. We read the data and find it. The material was real; the theme was assembled by expectation.
Protection: negative case analysis. Go find the people who do not fit. And check what each comment is actually *about*, not who wrote it.

2. Generalizing beyond our data.
Five people describe something, and we write "this group experiences X."
Protection: language discipline. Write **"these participants described"**, never **"this group experiences."** It is a small change, and it is the whole difference between a defensible finding and an indefensible one.

A warning that is worth more than the rule. When a barrier looks like it belongs to a cultural or language group, check whether it actually belongs to a **structure**: how people are employed, how they are paid, whether they have records, where they live. A finding that names the wrong cause does not just waste money. It confirms a story about a community that was never true, and it leaves the real barrier standing.

3. Treating one voice as the group's voice.
One powerful account becomes "the theme."
Protection: say plainly when we have one or two voices rather than a pattern.

The Thing People Get Wrong About Number 3

"Not a theme" does not mean "not important."

A single account can be the most action-worthy thing in the whole dataset, and still not be a finding. **Both of those are true at the same time.**

A finding is a claim about a pattern. One person is not a pattern. But one person can still be describing a real defect that will keep harming people, one at a time, until somebody fixes it.

Telling the difference between a finding and a story worth escalating is the actual skill. Not one or the other. Both, and knowing which is which.

If you leave this course believing that outlier voices do not count, you have learned the opposite of what this method is for.

What This Method Gives Us, and What It Does Not

It gives us: defensible descriptions of how different people experience our service. Enough to inform a Customer Experience Improvement Plan. Enough to justify a fuller study.

It does not give us: statistics, causes, prevalence estimates, or claims about anyone we did not hear from.

Those limits are features, not weaknesses. The honest analyst names them out loud. Naming them is what makes everything else we said believable.

What the Research Says, and Why We Do Not Just Cite the Number

You will find numbers in the literature. Twelve interviews. Nine. Sixteen to twenty-four. Twenty to forty. **They disagree with each other, and they are all correct**, because each one is the answer for *that* group asking *that* question.

Guest, Bunce and Johnson (2006) analyzed 60 in-depth interviews with women in two West African countries, on a narrow health topic. They read in batches of six and counted what each batch added.

After twelve interviews they had 100 *codes*, the labels a researcher attaches to pieces of text. That was 92% of the 109 that eventually came out of the thirty Ghanaian transcripts, and 88% of the 114 across both countries and all sixty interviews. Most of what the second country added was not new in substance. It was variation on themes already found.

They contributed the method: *read in batches, count what is new, watch the curve flatten.* That travels. The twelve does not, and the authors said so themselves. Their group was similar, their topic was narrow, their interview guide was structured. All three speed saturation up.

Hagaman and Wutich (2017) went back to that same study and tested it across multiple sites and cultures. Themes that cut across groups needed 20 to 40 interviews, not twelve.

Hennink, Kaiser and Marconi (2017) split saturation in two:

- **Code saturation**, no new *labels*, arrives around nine interviews.
- **Meaning saturation**, no new *understanding*, takes 16 to 24.

We will hit the first long before the second. When we stop hearing new labels, we are not finished. We have stopped learning what to *call* things. We have not stopped learning what they *mean*.

Guest, Namey and Chen (2020) give a practical, reportable procedure for assessing saturation as you go. If you want the formal version of the Hold-Back Test, this is it.

The point of all this: the literature tells us saturation is real, that it is reachable, and that the number depends. It does not tell us our number. **Only our data can do that, and only if we test it.**

Sources

- Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101.
- Guest, G., Bunce, A. and Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field Methods*, 18(1), 59–82.
- Guest, G., Namey, E. and Chen, M. (2020). A simple method to assess and report thematic saturation in qualitative research. *PLOS ONE*, 15(5).
- Hagaman, A. and Wutich, A. (2017). How many interviews are enough to identify metathemes in multisited and cross-cultural research? *Field Methods*, 29(1), 23–41.
- Hennink, M., Kaiser, B. and Marconi, V. (2017). Code saturation versus meaning saturation: how many interviews are enough? *Qualitative Health Research*, 27(4), 591–608.
- Lincoln, Y. and Guba, E. (1985). *Naturalistic Inquiry.* Sage.
- Shelton, R., Philbin, M. and Ramanadhan, S. (2021). Qualitative research methods in chronic disease and health equity. *Annual Review of Public Health.*

Ownership and Permitted Use

What you write here is yours. The document itself is ours. **So please do not upload this file, or any part of its content, into an AI tool, an AI-assisted learning management system, or any other system that stores or trains on what it receives.** That restriction is about protecting our work, not about limiting yours. If you want to think through your own entries with an AI tool, please copy your own answers out of the document and work with those.

Follow your agency AI use policy to determine if your answers and data may or may not be used with an AI tool.

You may not sell this document, use it to deliver training, or remove this notice. For any other use, reach out to us. Your facilitator: Carriann Lane at carriann.lane@sentient-learning.com.

Providing a copy on request, including under a public records request, does not grant any license to reuse.
