

AGENCY 150 | WHITE PAPER

When Fewer Interviews Beat More Surveys

Rethinking Statistical Power for Vega-Conducted Qualitative Research

May 2026

Lamar Pierce & Natasha Swift
Agency 150, Inc.

Executive Summary

A persistent critique levied against qualitative research methodologies is that they cannot match the statistical power of large-N quantitative surveys. The argument is familiar: quantitative work gives you the N you need to detect an effect; qualitative work gives you stories. This white paper argues that the critique is built on an unexamined assumption.

Statistical power is a function of two things, not one: sample size and measurement reliability. Standard power calculations cited in defense of large-N surveys silently assume a specific level of measurement noise — typically the noise of a single self-reported Likert item, with reliability coefficients often in the 0.6 to 0.7 range. Once measurement reliability is treated as a variable rather than a constant, the picture changes substantially.

Agency 150 conducts qualitative interviews through Vega, our AI-native interviewing platform. Vega runs structured 15-to-20-minute interviews against a literature-anchored protocol, in either voice or text modality, and produces transcripts that we score against validated coding rubrics. The combination of AI-conducted interviewing and a calibrated, human-supervised coding pipeline matters to this paper's argument in three ways. First, it reduces social-desirability bias on sensitive constructs — a result with strong empirical backing in the human-computer interaction literature. Second, it produces consistent depth and follow-up across respondents, which a fatigued or inconsistent human interviewer cannot match. Third, it makes high-reliability measurement cheap enough that the trade-off against survey N becomes a real choice rather than a theoretical one.

When a Vega interview is processed against a validated coding rubric, scored by calibrated coders (LLM-assisted under human supervision), and accompanied by a verbatim warrant quote for every score, the resulting measurement can achieve reliability in the 0.85 to 0.95 range. That higher reliability translates directly into

higher effective sample size: a study with N Vega-conducted interviews can carry the same statistical information as a substantially larger survey study using the same construct. This is not a workaround — it is the correct application of classical measurement theory, formalized by Spearman more than a century ago and reaffirmed in the contemporary methodological literature.

The implication for research buyers is consequential. If a Vega-derived score is more reliable than a self-report, more auditable at the row level, and grounded in validated scientific instruments, then the question is no longer "how many people do we need to survey" but "how much information does each observation carry, and what is the minimum N to reach the confidence threshold."

This paper makes that argument in six steps: it reframes the question; documents why self-reports are noisier than commonly assumed; presents the Vega-derived score as a validated measurement rather than a softer judgment; examines what AI interviewing specifically contributes (and where it has limits); treats the verbatim quote as the row-level validity warrant; and shows the math that links reliability to required sample size. We close with a decision framework, a candid statement of what we are and are not claiming, and the methodological caveats that govern the argument.

1. The Critique We Keep Hearing

The pro-quant objection comes in a recognizable form. A research team proposes a study based on 30 to 50 in-depth qualitative interviews. A reviewer or skeptical buyer responds with some version of: "We cannot make claims about a population from $N=30$. Our standard power calculation requires $N=400$ for a five-point margin at 95% confidence. Without that sample, this is exploratory at best." The argument has the appearance of statistical rigor and the standing of decades of textbook practice.

We want to take this objection seriously rather than dismiss it. The pro-quant camp is right about something real: sampling error scales with the square root of N , and a small qualitative sample cannot speak with the same precision about a population parameter as a large probability sample can. That part of the critique is correct, and this paper does not dispute it.

What we dispute is the implicit assumption that follows: that the only lever for confidence is sample size, and that every measurement strategy carries the same amount of noise per observation. Standard power formulas — the kind that produce "you need $N=400$ " — take measurement reliability as a fixed input. They quietly assume the reliability of a typical self-reported survey item, which is often weaker than the field acknowledges, and they offer no mechanism for crediting a more reliable measurement with the smaller N it actually requires.

The right question is not "how many people do you need to detect an effect?" The right question is: how much signal does each observation carry, and how many observations do you need to reach the confidence threshold? Sample size is one lever. Measurement reliability is the other. Power is the product, not the input.

When a vendor cites "you need $N=400$ for $\pm 5\%$," they are quoting a formula that assumes a specific level of measurement noise — usually the noise of a self-report Likert item. Change the measurement, change the math.

2. Why Self-Reports Are Noisier Than They Look

Before making the case for the Vega-derived score, we have to be honest about what the survey alternative actually delivers. The pro-quant default for measuring perceptions, attitudes, and experiences is the self-rated Likert scale: a respondent reads a statement and chooses a number from one to five or one to seven. This format is convenient, scalable, and familiar — and it is also subject to three well-documented sources of measurement error.

Construct ambiguity

A question like "How satisfied are you with this product?" means different things to different respondents. One person evaluates the product against a baseline of competitors; another against expectations set by marketing; another against an idealized version of the product they wish existed. The variance across responses reflects both real differences in satisfaction and differences in how each person interpreted the question. The resulting error term is invisible to the analyst and cannot be cleaned up after the fact.

Response-style bias

Respondents systematically differ in how they use rating scales independent of what they are rating. Some default toward agreement (acquiescence bias), some toward the middle of the scale (central-tendency bias), some toward the extremes (extreme-response bias), and some toward presenting themselves favorably (social-desirability bias). Podsakoff and colleagues, in the most-cited methodological review on this topic, document how these biases inflate measurement error and create spurious correlations between self-reported variables. None of these biases reflect the underlying construct — they reflect the format used to measure it.

Single-item unreliability

Most survey scores rest on one item, or a small handful. Classical test theory tells us that internal-consistency reliability — the share of observed variance that reflects true score variance rather than measurement error — increases with the number of items measuring the same construct. Cronbach's alpha for a single-item self-report measure rarely exceeds 0.7 and is frequently lower. This number matters considerably for power, and we will return to it in Section 6.

Cumulatively, these three sources of error mean a typical self-rated Likert item carries less signal than its tidy presentation suggests. The score of "4" entered into the dataset reflects some true underlying state, plus interpretation noise, plus response-style noise, plus single-item noise. The survey instrument cannot separate these. It can only collect them.

3. The Vega-Derived Score as a Validated Measurement

The pro-quant critique implicitly assumes that quantification is something only surveys do. The argument we want to make is the opposite: when a Vega interview is structured against a validated rubric and processed through a disciplined coding system, the resulting score is a measurement in the formal psychometric sense — and one with stronger construct-validity evidence than a typical self-report Likert item.

The case rests on three interlocking claims. First, construct validity in our codebooks comes from validated scientific instruments rather than being invented for the project. Second, that the operational coding process is itself a four-stage validation system, with humans doing the load-bearing methodological work and Vega plus our coding pipeline doing the bulk labor. Third, that the resulting score can be defended on both reliability and construct validity in ways that self-reports usually cannot.

3.1 Construct validity starts upstream: we do not invent the rubric

Construct validity asks whether a measure captures the underlying concept it claims to measure. It is the deepest psychometric property a measurement can have, and it is frequently the property that a self-report Likert scale gestures at rather than establishes. The pro-quant camp typically grants construct validity to surveys by reference — "we used a validated scale" — and denies it to qualitative work because qualitative work is interpretive.

Agency 150 codebooks are not built from scratch. They integrate published constructs and scoring anchors from peer-reviewed scales in the relevant scientific literature. The flagship anchor in our team and organizational research is Edmondson's seven-item team psychological safety scale, originally developed and validated in *Administrative Science Quarterly* in 1999. Beyond that, depending on the research question, our codebooks draw on instruments such as the Unified Theory of Acceptance and Use of Technology (UTAUT2) for technology adoption work; the McKnight et al. trust-in-technology framework when the object of study is an AI system or other software; the Mayer, Davis, and Schoorman ability-benevolence-integrity model when the construct is interpersonal or organizational trust; the Utrecht Work Engagement Scale (UWES) for engagement; and the Chung et al. Work Group Inclusion scale for inclusion and belonging.

We adapt the operationalization — the questions Vega asks, the prompts, the elicitation strategy — to the AI-conducted qualitative format. But the construct definitions and scoring anchors are inherited from instruments with documented psychometric properties. This matters for two reasons. First, it means the construct-validity claim is imported from peer-reviewed work; it is not invented at Agency 150. Second, it means the Vega-derived score is directly comparable to scores from prior literature using the same construct, which an ad hoc survey instrument typically cannot claim.

The codebook is built on scales scientists have already validated. We are not asking respondents to rate themselves on those scales — we are using those scales as the coder's yardstick against transcript evidence. That is a meaningfully stronger position than a self-report Likert.

3.2 The coding process is the warrant: human-validated, scaled by automation

A construct definition becomes a reliable measure only when applied consistently across cases. Agency 150 treats this as a four-stage validation system. As an AI-native research firm, we are explicit about where automation contributes and where human judgment is irreducible.

Stage 1 — Codebook authoring (human)

Researchers translate the literature-anchored constructs into the project-specific rubric: scoring anchors, decision rules, exclusion criteria, and worked examples for edge cases. They also design the Vega interview protocol — the questions, follow-up logic, and elicitation strategy — that will draw out the evidence the codebook needs to score against. This stage is irreducibly human. It encodes tacit construct knowledge and the specific framing the research question demands. An LLM can draft a starting codebook from a research brief and produce something plausible, but the iteration of "no, that is not what we mean by usefulness in this context — here is why" requires a human researcher who understands both the construct and the domain.

Stage 2 — Calibration on a gold-standard set (human)

Before any scoring is done at scale, two analysts independently code a calibration sample of Vega transcripts against the rubric. Inter-rater agreement is measured using established metrics — typically Cohen's kappa for categorical codes or Krippendorff's alpha when codes vary in level of measurement. Krippendorff has set the standard threshold for reliable coding at $\alpha \geq 0.80$, with $\alpha \geq 0.67$ acceptable for

tentative conclusions; in clinical and high-stakes domains, $\alpha \geq 0.90$ is often required. Disagreements on the calibration set are adjudicated, and the rubric is refined until agreement clears the threshold. This is the qualitative equivalent of pilot-testing a survey instrument, and it produces a documented reliability statistic before the main study begins.

Stage 3 — Scaled coding (LLM-assisted, human-supervised)

Once the rubric is calibrated, the bulk of transcript-by-transcript scoring is done with LLM-assisted coding, prompted against the locked rubric and required to extract a verbatim warrant quote for every score it assigns. This is where the cost and speed advantage over surveys actually materializes — a 15-to-20-minute Vega transcript can be coded against a 12-construct rubric in under a minute. Recent peer-reviewed work has documented that LLMs operating against well-specified codebooks can match or exceed human inter-rater reliability on many latent content-analysis tasks, with Cohen's kappa frequently exceeding 0.7 in deductive coding settings, and that LLM raters show very high temporal consistency across repeated coding passes.

We are explicit about this. The LLM is not the methodological authority; it is executing a protocol that humans authored, calibrated, and continue to supervise. But the labor it performs is real, and the cost structure it produces is what makes the argument in Section 6 — that Vega-derived scores can be operationally competitive with surveys at scale — a practical claim and not just a methodological one.

Stage 4 — Adjudication and spot-audit (human)

The system would not be defensible without the closing human review. Cases the model flags as low-confidence, cases where two independent coding passes disagree, and a random ~5% sample of the final dataset are reviewed by a human analyst. This catches drift, surfaces edge cases the calibration set missed, verifies that warrant quotes are not confabulated, and produces the audit trail that makes the resulting dataset defensible to a sophisticated buyer.

The full chain — literature-grounded constructs, human-authored rubric and Vega protocol, human-calibrated reliability, automation-scaled coding with required verbatim warrants, and human adjudication and spot-audit — is what makes the resulting numbers a measurement rather than an opinion. Humans move from coding 400 transcripts to validating a system that codes 400 transcripts. The methodological footprint is the same; the operational footprint is dramatically smaller.

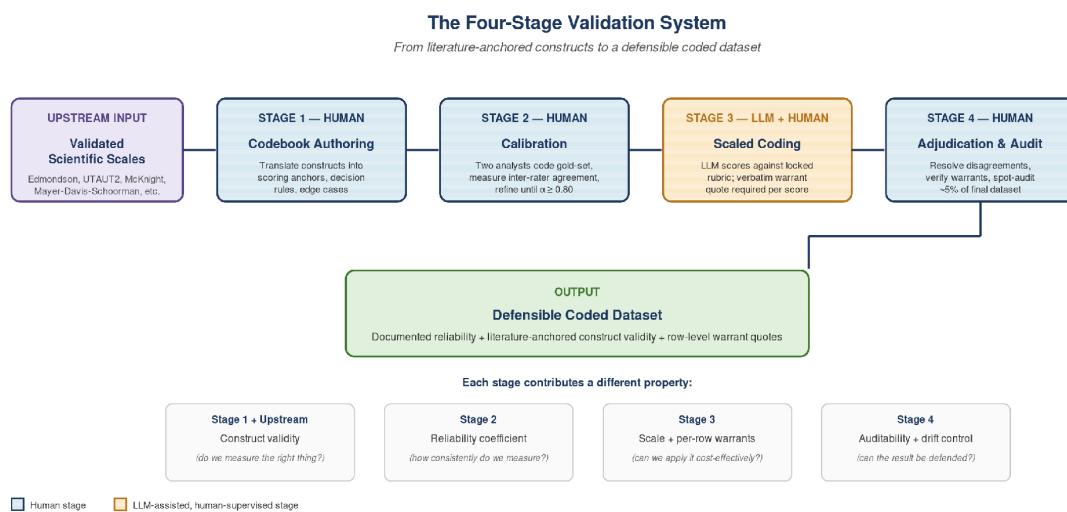


Figure 1. The four-stage validation system, with upstream construct inputs and the property each stage contributes to the final dataset.

3.3 Why this beats the survey on validity, not just reliability

Reliability and construct validity are different properties. A measure can be reliable without being valid (consistently measuring the wrong thing) or valid without being reliable (loosely capturing the right thing). The Vega-derived score can be defended on both, and that twin defense is what gives the argument in Section 6 its weight.

Reliability comes from the calibrated coding protocol. Documented inter-rater agreement on the calibration set yields a published reliability coefficient — typically 0.85 to 0.95 in well-run two-coder adjudication, and comparable when the second

coder is the LLM operating against the locked rubric and verified by spot-audit. This number goes directly into the power calculation in Section 6.

Construct validity comes from three sources at once: the constructs are inherited from validated scientific scales rather than authored ad hoc; the coder is rating transcript behavior and content rather than asking the respondent to self-locate on an abstract scale; and every score carries a verbatim warrant — which is itself a piece of construct-validity evidence at the row level (Section 5).

The contrast with a typical self-report survey is sharp. A self-rated 4 on "I find this useful" depends on how the respondent interprets useful, what reference point they are comparing against, what mood they are in, and what response style they default to. A coded 4 against a literature-anchored rubric, with a quote attached, depends on whether the transcript actually exhibits the behaviors the construct specifies — which is, by definition, what the construct claims to measure.

4. What Vega Specifically Contributes (and Where It Has Limits)

The argument in Section 3 holds for any qualitative interview processed through a calibrated coding pipeline. But Vega — Agency 150's AI-native interviewing platform — is doing additional work beyond what a traditional human-interviewer protocol does, and being honest about what that work is (and where its limits sit) sharpens the case.

Vega conducts structured 15-to-20-minute interviews in voice or text, using a research-team-authored protocol with adaptive follow-up logic—three properties of this setup matter for the measurement argument.

4.1 AI interviewers reduce social-desirability bias on sensitive constructs

A well-replicated finding in the human-computer interaction literature is that respondents disclose more honestly to a computer interviewer than to a human one,

particularly on topics where the respondent has reason to manage their self-presentation. Lucas and colleagues at USC's Institute for Creative Technologies showed in a foundational 2014 study that respondents who believed they were speaking with an automated virtual interviewer were significantly more willing to disclose stigmatized information than respondents who believed they were speaking with a human. Subsequent work in military mental-health screening showed that returning service members reported more depression and PTSD symptoms during virtual-interviewer conversations than on the official paper-based PDHA. The effect is consistent and large enough to be meaningful for measurement.

The mechanism is straightforward: respondents perceive less judgment from a non-human interlocutor, which lowers the impression-management overhead they would otherwise apply to their answers. The effect is most pronounced on constructs where social desirability is high — workplace conflict, perceived fairness, mental health, performance shortfalls, fairness of treatment by managers, attitudes toward colleagues, ethically ambiguous behaviors, and similar topics. This is also part of the broader literature on computerized self-administration: a meta-analysis covering more than six hundred effect sizes found that computer-administered measures produce less social-desirability distortion than paper or face-to-face equivalents, especially when respondents are alone and can backtrack.

For the measurement argument in this paper, the implication is that Vega is not just a more efficient interviewer — for sensitive constructs, it is a more accurate one. The reliability number used in the power calculation in Section 6 reflects not only the consistency of the coding pipeline but also the honesty of the underlying disclosure. A self-report survey on a sensitive topic carries the social-desirability error term whether the respondent fills it out alone or not; a Vega interview reduces that term at the source.

Respondents disclose more honestly to a computer interviewer than to a human one. This is not a soft claim. It is one of the most replicated findings in the human-computer interaction literature.

4.2 Vega delivers consistent depth across respondents

A second property of AI-conducted interviewing is that the interviewer does not get tired, distracted, or rapport-blocked late in the session. A human interviewer running their fifth interview of the day on a difficult topic will, even with the best training, ask shorter follow-ups, accept thinner answers, and miss inconsistencies more often than they did in interview one. The variance across interviews this introduces is silent measurement error: respondent four's coded score reflects partly her experience and partly the fact that her interviewer was sharper than respondent fifteen's.

Vega holds the interview protocol constant. The same construct gets the same elicitation strategy on respondent one and respondent five hundred. Adaptive follow-ups are triggered by what the respondent says, not by how energized the interviewer is feeling. This consistency does not eliminate respondent-level variance — different respondents will still give different amounts of detail — but it eliminates an entire layer of interviewer-driven variance that the standard qualitative methodology either lives with or expensively controls for through extensive interviewer training.

There are constructs where a human interviewer's accumulated rapport produces depth Vega may not match. Long, narrative life-history interviews, deeply emotional disclosures that benefit from human warmth, and topics where the respondent needs to feel heard before they will speak honestly are all candidates. We are candid about this: Vega is not a universal substitute for skilled human qualitative interviewing. For the bounded, construct-focused, score-producing work this paper is about — measuring psychological safety, trust, perceived usefulness, fairness, and

similar constructs in 15-to-20-minute sessions — Vega's consistency advantage is, in our judgment, the dominant effect. For deeper ethnographic or narrative work, the trade-off shifts.

4.3 Voice versus text: an open empirical question

Vega supports both voice (spoken response, transcribed) and text (typed response) modalities. The honest answer about which produces better data for measurement-grade qualitative work is that we do not yet have a settled view, and the published literature does not provide one either.

What the evidence does say: a 2024 study by Höhne and colleagues comparing voice and text answers to open-ended sensitive questions in smartphone surveys found that voice answers were longer in word count and covered more topics than text answers, but item nonresponse was substantially higher for voice (25 to 60 percent) than for text (around 5 percent or less). Sentiment and answer extremity did not differ significantly. Read together, this suggests voice produces richer content from the respondents who engage, but a meaningful share of respondents simply will not answer at all in voice mode — they skip the question.

There are plausible directional arguments on each side that the existing literature does not yet resolve. Voice may surface affect, hesitation, and emphasis that text strips away — features that a sophisticated coding protocol could (in principle) score against. Text may reduce the social-pressure component of voice by giving respondents time to consider their answer, re-read, and edit; this could reduce social-desirability bias further than voice does, since audio recording feels more like being heard than typing does. Different respondents likely prefer different modalities for non-content reasons (typing speed, accent self-consciousness, environment), and forcing one modality on all respondents introduces selection effects that bias the sample.

Our current operational stance reflects this uncertainty. Vega offers respondents the choice of modality where the research question allows, and we treat modality as a covariate to monitor in analysis rather than a settled methodological choice. Agency 150 is collecting evidence on this question across our active studies, and we expect to report findings in subsequent work. For the measurement argument in this paper, the relevant point is that the construct-validity and reliability claims in Section 3 hold for both modalities — a coded score against a calibrated rubric is a coded score against a calibrated rubric, whether the underlying transcript came from voice or text. Modality may shift the level of the reliability coefficient up or down at the margin; it does not change the structure of the argument.

5. The Quote as Row-Level Validity Warrant

The verbatim quote attached to every coded score is not an aesthetic flourish or a presentation device. It is the structural feature that makes the per-observation validity claim auditable, and it is the feature that keeps automated coding honest when the coder is an LLM rather than a human.

Every Vega-derived score travels with its evidence. A row in the dataset reads, in effect: "We coded this respondent as a 4 on usefulness because she said: '[quote].'" This does two things at once.

First, it makes the score externally auditable. A survey score of 4 is unfalsifiable. No one can test whether the respondent's self-rating honestly reflects their experience or is driven by acquiescence, fatigue, or misinterpretation. A coded score of 4 with a Vega transcript and a warrant quote is auditable by every one of those parties — and when the coder is automated, also by the human spot-auditor verifying the model did not confabulate the quote.

Second, it provides construct-validity evidence at the row level. In a survey, construct validity is established (if at all) once, at the instrument level, in the abstract — a

scale was validated in a different population, perhaps decades ago, and the validity claim travels by reference. In a quote-warranted coded dataset, validity is re-established for every single observation. The buyer can read the quote and judge whether the score reflects the construct. So can a peer reviewer. So can the analyst's future self.

The implication for AI-native coding is meaningful. The "robot coder" critique a sophisticated buyer might raise — that an LLM will confidently assign a code to anything you give it, even when the construct is not present — is structurally answered by the warrant requirement. A score without an in-transcript quote that supports it cannot be entered. A score with a confabulated quote is detectable on spot-audit. The warrant requirement, in effect, converts the LLM from a confident generator of plausible labels into a constrained executor of a validated protocol.

A survey score of 4 is unfalsifiable. A coded score of 4 with a Vega quote attached is auditable by the client, by a second coder, and by the analyst's future self.

6. The Math: Reliability Shrinks the N You Need

We promised the math would be conceptual rather than dense. The load-bearing relationship is straightforward, and once it is on the page, the rest follows.

In his 1904 paper on the proof and measurement of association, Spearman derived the formula for how measurement error attenuates an observed correlation. The contemporary application of that result to power and sample size is direct. When a measure has reliability less than 1.0, the true effect size is attenuated in the data, and detecting the same underlying effect requires more observations to compensate. The relationship, in plain form:

Effective sample size $\approx N \times \text{reliability (of the dependent measure)}$

A study of $N=400$ respondents using a measure with reliability 0.6 carries roughly the same statistical information as a study of $N=240$ respondents using a measure with reliability 1.0. This simplifies the noncentrality-parameter relationship that governs t-tests, regression, and correlation tests; the exact form depends on the specific test, but the direction of the relationship is robust across standard inferential machinery. Recent methodological work has formalized this into adjusted sample-size formulas that explicitly include reliability as a parameter — Charles's 2018 modification of Spearman's attenuation equation, for instance, provides partial-correction sample-size calculations that practitioners can apply directly. The full derivation is provided in Appendix A.

To make this concrete, the table below shows the required sample size to detect a small-to-medium effect (Cohen's $d = 0.3$) at 80% power and $\alpha = 0.05$ across measurement strategies with varying reliability. The reliability values are realistic targets for each measurement type based on the literature reviewed in Sections 2 and 3.

Measurement strategy	Typical reliability	Effective N at N=300	N to detect $d=0.3$
Single self-report Likert item	0.60	180	~290
Multi-item validated survey scale	0.80	240	~220
Coded Vega interview score	0.90	270	~195
Double-coded Vega with adjudication	0.95	285	~185

Estimates assume a two-tailed independent-samples test, $\alpha = 0.05$, target power = 0.80, and a true population effect of $d = 0.3$. Required N is computed by inflating the

unattenuated requirement by $1/\text{reliability}$ following Spearman's attenuation logic; results align with standard reliability-adjusted power tables.

The point of the table is not that qualitative work always wins. It is that any honest power calculation must include a reliability term. Once it does, the gap between "the N you need for a survey" and "the N you need for a Vega-derived score" closes — and often crosses over. A study using a calibrated Vega protocol with reliability 0.90 needs roughly two-thirds the sample size of a single-item self-report study to detect the same effect at the same confidence level. Adjudicated, double-coded Vega work tightens that ratio further.

The saturation literature in qualitative methodology has documented an additional consideration for two decades. In their landmark 2006 study, Guest, Bunce, and Johnson found that thematic saturation in semi-structured interviews was reached within the first twelve interviews, with most metathemes present by six. Hennink and colleagues subsequently distinguished code saturation (around nine interviews) from meaning saturation (sixteen to twenty-four), and Malterud and colleagues reframed the question entirely with the concept of information power: the more information each unit of data carries relevant to the study question, the smaller the sample needed to support a defensible inference. Taken together, these literatures establish that qualitative sample sizes do not need to be large to be informative; the question is whether the per-respondent information yield justifies the chosen N.

What we are arguing is that the same logic applies, with sharper teeth, to Vega-conducted interviews scored into quantitative measurements. The information yield per respondent is higher than a self-report because the measurement is more reliable; therefore, the N required for a given confidence is lower; and therefore, the standard "you need 400" critique is using the wrong yardstick.

7. A Decision Framework

Even though our central claim favors Vega-derived scores, we want to give readers a framework rather than a verdict. There are research questions where surveys remain the right tool. The framework below sketches when each method is the better fit.

- **Reach for a large-N survey when:**
 - The construct is well-served by an existing single-item or short-scale self-report (e.g., demographic facts, simple behavioral frequencies, well-validated short attitude scales).
 - The research question is about population-level prevalence with tight precision requirements — for example, "What share of customers prefer X?" with a margin of error under three points.
 - Cost or access constraints make 15-to-20-minute Vega interviews impractical, and the construct can tolerate the measurement noise of self-report.
- **Reach for a Vega-conducted study when:**
 - The construct is complex, multi-faceted, or context-dependent in ways a Likert item compresses badly (psychological safety, trust, perceived fairness, quality of judgment).
 - The research question is about a sensitive topic where social-desirability bias would distort self-reports — workplace conflict, fairness, mental health, ethically ambiguous behavior, attitudes toward managers or colleagues.
 - The research question requires defending each observation, not just the aggregate — for example, in regulated contexts, executive briefings, or scientific publication where construct validity is contested.

- The downstream use of the data benefits from row-level interpretive evidence, such as quote selection for case studies or qualitative substantiation of quantitative claims.
- Sample access is constrained — the population is small, hard to reach, or expensive — and per-respondent information yield is the binding constraint rather than total N.

A useful diagnostic question for the buyer: "If the headline finding is challenged, what evidence do I have?" If the answer is "an aggregate from N=400 self-reports," the construct better be unambiguous, and the question better be one of prevalence. If the answer is "a defensible measurement on each respondent, with the verbatim Vega evidence behind every score," the qualitative-derived approach is the more rigorous instrument, even at substantially lower N.

8. What This Paper Is Not Claiming

To preempt the pushback that will come from rigorous quantitative readers, we want to be explicit about the boundaries of the argument.

- We are not claiming that qualitative research replaces survey work for every research question. The decision framework in Section 7 is a frank attempt to mark where each tool is appropriate.
- We are not claiming that small-N qualitative studies achieve the same population generalizability as large-N probability surveys. Power and generalizability are distinct properties; this paper is about the first. A Vega-derived score with high reliability still depends on a thoughtful sampling strategy for its population claims.
- We are not claiming that all qualitative interviews produce reliable scores. The argument depends on structured Vega protocols, literature-anchored codebooks, calibrated coders, documented inter-rater reliability, and human

supervision of automated coding. Without those, qualitative scores can be less reliable than self-reports, not more. The methodology is the warrant; remove it, and the argument collapses.

- We are not claiming that AI interviewing universally outperforms skilled human qualitative work. For deep narrative ethnography, life-history interviews, and emotionally complex disclosures that benefit from accumulated human rapport, a trained human interviewer may produce richer data than Vega does. The argument here is bounded to the construct-focused, score-producing measurement work for which Vega is designed.
- We are not claiming a settled view on voice versus text modality. The published literature does not yet support a strong directional claim, and Agency 150 is actively gathering evidence. Section 4.3 frames this as the open question it currently is.
- We are not claiming that sampling error disappears. It still scales with the square root of N . The argument is that measurement error — a separate, multiplicative source of noise — is the unaccounted-for term in the standard "you need $N=400$ " critique. Once both terms are in the equation, the trade-off becomes negotiable rather than foreordained.
- We are not claiming that LLM-assisted coding is a substitute for methodological judgment. The four-stage validation system in Section 3.2 puts humans on the load-bearing methodological work — codebook authoring, calibration, adjudication, and spot-audit — and uses automation to scale the bulk labor. Removing the human stages voids the validity claim.

9. Conclusion

Statistical power is the product of sample size and measurement reliability, not sample size alone. The pro-quant critique of qualitative research has, for too long,

treated reliability as a fixed background condition and arrived at "you need a larger N" as a foregone conclusion. Once reliability is treated as a variable — and once measurement strategies are evaluated on what they actually contribute to it — the conclusion changes.

A Vega interview structured around a literature-anchored rubric, scored with a calibrated coding system requiring verbatim warrants, and supervised through human adjudication and spot-audits produces a measurement that can be more reliable, more construct-valid, and more auditable than a typical self-report survey. Vega-conducted interviewing adds two specific advantages on top of that: reduced social-desirability bias on sensitive constructs, with strong empirical backing in the human-computer interaction literature; and consistent depth of elicitation across respondents that human interviewers cannot match at scale. The math, applied honestly, says the Vega-derived approach can reach the same statistical confidence at materially lower N. The methodology, applied rigorously, says the resulting numbers are a measurement in the formal psychometric sense, not a softer judgment in quantitative dress.

For research buyers weighing the next study, the question worth asking is not "how many people do we need to survey." It is: "What is the most reliable measurement I can build for this construct, what does that reliability buy me in effective sample size, and what evidence do I want to be able to put behind every observation when the headline finding is questioned?" That reframing is the contribution of this paper. The methodology, the literature, and the math are all available to support it.

Appendix A. Formal Derivation of Reliability-Adjusted Sample Size

This appendix gives the explicit math behind the central claim of Section 6: that required sample size for a given level of statistical power is inversely proportional to the reliability of the dependent measure. The treatment follows classical test theory, Spearman's 1904 attenuation result, and the noncentrality-parameter logic that underlies modern power calculations. It is presented for readers who want to verify the argument rather than take it on conceptual faith.

A.1 Classical test theory decomposition

Classical test theory models any observed score X as the sum of a true score T and a measurement-error term e , with the standard assumptions that the error has mean zero and is uncorrelated with the true score:

$$X = T + e, \quad E[e] = 0, \quad \text{Cov}(T, e) = 0 \quad (\text{A.1})$$

Under this model, the variance of the observed score decomposes additively into true-score variance and error variance:

$$\text{Var}(X) = \text{Var}(T) + \text{Var}(e) \quad (\text{A.2})$$

The reliability coefficient ρ_{XX} is defined as the proportion of observed-score variance that is true-score variance:

$$\rho_{XX} = \text{Var}(T) / \text{Var}(X) = \text{Var}(T) / [\text{Var}(T) + \text{Var}(e)] \quad (\text{A.3})$$

Reliability is bounded between 0 and 1. $\rho_{XX} = 1$ corresponds to a noiseless measurement; $\rho_{XX} = 0$ corresponds to a measurement that is pure error. For a self-rated single Likert item, published estimates typically place ρ_{XX} in the 0.6 to 0.7 range; for a calibrated multi-item scale, 0.8 is a common target; for a Vega-derived score with documented inter-rater agreement, 0.85 to 0.95 is achievable. Cronbach's alpha is one estimator of ρ_{XX} for multi-item scales; Krippendorff's alpha and Cohen's kappa serve the same role for coded data.

A.2 Spearman's attenuation result

Spearman (1904) derived the relationship between an observed correlation between two measures and the correlation between their underlying true scores. If r_{XY} is the observed correlation and $\rho_{TX, TY}$ is the true-score correlation, and ρ_{XX} and ρ_{YY} are the reliabilities of X and Y respectively, then:

$$r_{XY} = \rho_{TX, TY} \cdot \sqrt{(\rho_{XX} \cdot \rho_{YY})} \quad (\text{A.4})$$

The observed correlation is the true correlation attenuated by the geometric mean of the two reliabilities. When only the dependent measure has appreciable measurement error, and the predictor (e.g., a treatment indicator or demographic variable) is measured without error, $\rho_{XX} = 1$ and (A.4) simplifies to:

$$r_{XY} = \rho_{TX, TY} \cdot \sqrt{(\rho_{YY})} \quad (\text{A.5})$$

The attenuation result shows that measurement error in the dependent variable systematically shrinks the observed effect, regardless of the true effect size. This is the source of the power loss the present argument addresses.

A.3 From attenuation to the noncentrality parameter

The statistical power of a hypothesis test depends on the test statistic's noncentrality parameter (NCP) under the alternative hypothesis. For the two-sample t-test, the NCP is:

$$\lambda = (\mu_1 - \mu_2) / \sigma \cdot \sqrt{(n/2)} \quad \text{for equal } n \text{ per group} \quad (\text{A.6})$$

where $(\mu_1 - \mu_2) / \sigma$ is the standardized effect size in true-score units, and n is the per-group sample size. When the dependent variable is measured with reliability ρ_{YY} , the observed standard deviation σ_Y is inflated by the measurement error:

$$\sigma_Y = \sigma_T / \sqrt{(\rho_{YY})} \quad (\text{A.7})$$

which follows directly from (A.2) and (A.3). Substituting into the standardized effect size, the observed effect size in the data is:

$$d_{obs} = d_{true} \cdot \sqrt{(\rho_{YY})} \text{ (A.8)}$$

and the realized noncentrality becomes:

$$\lambda_{obs} = d_{true} \cdot \sqrt{(\rho_{YY})} \cdot \sqrt{(n/2)} \text{ (A.9)}$$

To achieve a target NCP λ^* (which determines power for a given α), the required sample size scales as:

$$n_{required} \propto 1 / (d_{true}^2 \cdot \rho_{YY}) \text{ (A.10)}$$

Equation (A.10) is the load-bearing result. For a fixed target effect size and power, the required sample size is inversely proportional to the dependent measure's reliability. This is the formal version of the conceptual statement in Section 6 that "effective sample size $\approx N \times$ reliability."

A.4 The reliability-adjusted N formula used in Table 1

Let n_0 denote the sample size required to achieve target power $1 - \beta$ under perfect reliability ($\rho_{YY} = 1$) for a given true effect size d , significance level α , and test type. Standard sample-size tables (e.g., Cohen, 1988) and statistical software (G*Power, R's pwr package) report n_0 directly. The reliability-adjusted requirement under measurement reliability ρ_{YY} is:

$$n(\rho_{YY}) = n_0 / \rho_{YY} \text{ (A.11)}$$

For the table in Section 6, we set $d = 0.3$, $\alpha = 0.05$ (two-tailed), and target power = 0.80, which yields $n_0 \approx 175$ per group from standard tables, or roughly 175 total in a one-sample / paired-difference framing. Applying (A.11) to the four reliability levels in the table:

Measurement strategy	ρ_{YY}	n_0	$1 / \rho_{YY}$	$n(\rho_{YY})$
Single self-report Likert	0.60	175	1.667	292
Multi-item validated scale	0.80	175	1.250	219
Coded Vega score	0.90	175	1.111	195
Double-coded with adjudication	0.95	175	1.053	184

These are the values reported in Section 6 (rounded to the nearest 5). The same logic applies to other tests with adjustments for the relevant NCP form: for Pearson correlation, n scales as $1/\rho_{YY}$ in the same way; for multiple regression, it is the reliability of the criterion that drives the adjustment, with predictor reliabilities affecting standard errors of individual coefficients through (A.4).

A.5 Notes, limits, and partial corrections

Three caveats accompany this derivation, each of which a methodologically sophisticated reader will want noted.

First, equation (A.11) treats reliability as a known constant. In practice, ρ_{YY} is estimated from data (Cronbach's alpha for multi-item scales; Krippendorff's alpha or Cohen's kappa for coded data), and the estimate has its own sampling variability. Charles (2018) develops partial corrections that propagate this uncertainty into the sample-size calculation; the practical effect is to shift required n upward modestly when the reliability estimate is itself imprecise. We recommend his procedure for studies where the reliability estimate is based on a small calibration set.

Second, the simple inverse-reliability scaling in (A.11) holds exactly when the predictor is measured without error. When both X and Y have measurement error, the geometric-mean form (A.4) applies, and the effective- N adjustment is $\sqrt{(\rho_{XX} \cdot$

ρ_{YY}) on the effect size, giving $n_{\text{required}} \propto 1/(\rho_{XX} \cdot \rho_{YY})$. For most field studies where the predictor is a treatment indicator or demographic variable (treated as error-free), the (A.11) form is appropriate.

Third, the reliability of a Vega-derived score is itself a measurement that requires defending. Reporting Krippendorff's alpha or Cohen's kappa from the calibration set, with the size of that calibration set, is the minimum disclosure that allows a reviewer to assess whether the reliability number entering (A.11) is itself reliable. We recommend a calibration set of at least 20 to 40 transcripts independently coded by two analysts, with disagreements adjudicated and the rubric refined until $\alpha \geq 0.80$ (or the discipline-appropriate threshold) before scaled coding begins.

None of these caveats change the central conclusion of Section 6. They sharpen the conditions under which it applies. The proposition that "required N is inversely proportional to dependent-measure reliability" is a hundred-and-twenty-year-old result in measurement theory. This paper's contribution is methodological rather than mathematical: applying that result to the choice between Vega-conducted qualitative measurement and self-reported quantitative scores, with explicit attention to the validation system that produces the reliability number on the right-hand side of the equation.

References

All references below have been verified against primary sources accessible at the URLs cited.

Charles, E. P. (2018). Modifying Spearman's attenuation equation to yield partial corrections for measurement error — with application to sample size calculations. [Educational and Psychological Measurement](#), 78(1), 100–114.

- Chung, B. G., Ehrhart, K. H., Shore, L. M., Randel, A. E., Dean, M. A., & Kedharnath, U. (2020). Work group inclusion: Test of a scale and model. [Group & Organization Management](#), 45(1), 75–102.
- Cohen, J. (1988). [Statistical Power Analysis for the Behavioral Sciences](#) (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cronbach, L. J. (1951). [Coefficient alpha and the internal structure of tests](#). *Psychometrika*, 16(3), 297–334.
- Edmondson, A. C. (1999). [Psychological safety and learning behavior in work teams](#). *Administrative Science Quarterly*, 44(2), 350–383.
- Guest, G., Bunce, A., & Johnson, L. (2006). [How many interviews are enough? An experiment with data saturation and variability](#). *Field Methods*, 18(1), 59–82.
- Hennink, M. M., Kaiser, B. N., & Marconi, V. C. (2017). [Code saturation versus meaning saturation: How many interviews are enough?](#) *Qualitative Health Research*, 27(4), 591–608.
- Höhne, J. K., Gavras, K., & Claassen, J. (2024). [Typing or speaking? Comparing text and voice answers to open questions on sensitive topics in smartphone surveys](#). *Social Science Computer Review*.
- Krippendorff, K. (2011). [Computing Krippendorff's alpha-reliability](#). Departmental Papers (ASC), University of Pennsylvania.
- Lucas, G. M., Gratch, J., King, A., & Morency, L.-P. (2014). [It's only a computer: Virtual humans increase willingness to disclose](#). *Computers in Human Behavior*, 37, 94–100.
- Lucas, G. M., Rizzo, A., Gratch, J., Scherer, S., Stratou, G., Boberg, J., & Morency, L.-P. (2017). [Reporting mental health symptoms: Breaking down barriers to care with virtual human interviewers](#). *Frontiers in Robotics and AI*, 4, 51.

- Malterud, K., Siersma, V. D., & Guassora, A. D. (2016). [Sample size in qualitative interview studies: Guided by information power](#). *Qualitative Health Research*, 26(13), 1753–1760.
- Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). [An integrative model of organizational trust](#). *Academy of Management Review*, 20(3), 709–734.
- McKnight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). [Trust in a specific technology: An investigation of its components and measures](#). *ACM Transactions on Management Information Systems*, 2(2), Article 12.
- Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). [Common method biases in behavioral research: A critical review of the literature and recommended remedies](#). *Journal of Applied Psychology*, 88(5), 879–903.
- Richman, W. L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). [A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews](#). *Journal of Applied Psychology*, 84(5), 754–775.
- Schaufeli, W. B., & Bakker, A. B. (2003). [UWES — Utrecht Work Engagement Scale: Preliminary Manual](#). Occupational Health Psychology Unit, Utrecht University.
- Spearman, C. (1904). [The proof and measurement of association between two things](#). *American Journal of Psychology*, 15(1), 72–101.
- Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). [Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology](#). *MIS Quarterly*, 36(1), 157–178.
- Zhu, J., Li, Z., & Li, J. (2025). [Comparing large language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity, and sarcasm](#). *Scientific Reports*, 15.