Remote Moderated vs. Remote Automated – Different outcomes?

Some time ago I began investigating the data outcome differences between **remote *moderated*** research and **remote *automated*** research (the main service provided by UserZoom).

Surprisingly, there hasn't been a lot of scientifically controlled, empirical data around the feedback quality between the two different types of research methods. Mostly, the existing research tracks differences in usability issues found [1][2][3], but remote research methods have evolved far beyond merely assessing usability issues.

I wanted to examine a wider, holistic view of whether the two methods create different feedback across many variables. Ultimately, this exploration hoped to provide data so companies could better understand their qualitative research options based on goals rather than fear towards new methods.

Usability should only be one component of the larger need for feedback.

How We Conducted the Research…

- 40 participants for remote moderated
- 40 participants for remote automated

Participants were screened and recruited via YouEye's panel (now UserZoom's!) which has been built and grown over the previous 5 years. *Panel usage has huge implications on feedback outcome which will be discussed further below.*

Controlled for these main variables…

- Identical *pre-determined* prompts (moderator could still present follow-ups)
- Timing of stimulus exposure (after the proper tasks and intro)
- Identical solicited ratings (Likert, single-choice, etc.)
- Similar demographic distribution between the two testing groups.
  - ➢ Familiarity with testing method was balanced as well!

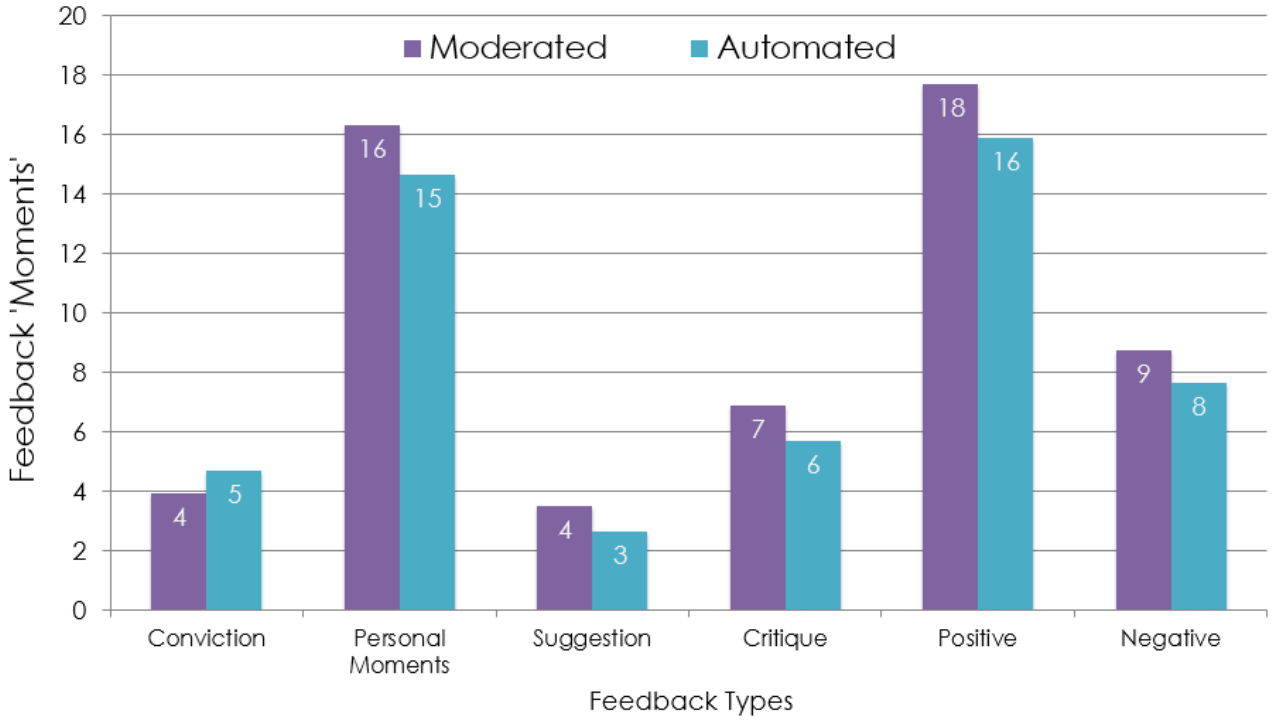
Simply put, the study design was near identical for the two experiences.
There ***are*** intrinsic differences with moderated sessions, and those are explained further below.

The subject matter of the study design was music listening habits and online music behavior.

What 'moments' we measured… [Click here to jump to measure definitions]
Qualitative (open-ended) feedback categorized as 'moments' into feedback themes…

- Conviction, Unique Supported Reasonings, & Personal Moments
- Suggestion, Critique, & Praise
- Minor, Major, & Critical usability issues identified
- Positive & Negative

Transcribed results were coded by an ***external*** (non YouEye nor UserZoom) researcher and the coding instructions were identical for all transcripts. Measures framework was created over 5 multi-hour sessions utilizing standard IRR practices.



Despite slight variances, none of these measures resulted in statistically significant differences.

Most Feedback Measures Were Equal

As you can see, the real goal was to explore areas of importance that matter to a researcher when it comes to *feedback outcome*. What we found across the board was that the quality and consistency of the feedback was not impacted in **nine out of eleven** metrics that we tracked. The results were really interesting, and have huge implications for remote automated research, as well as highlighting the *strengths of remote moderated* research.

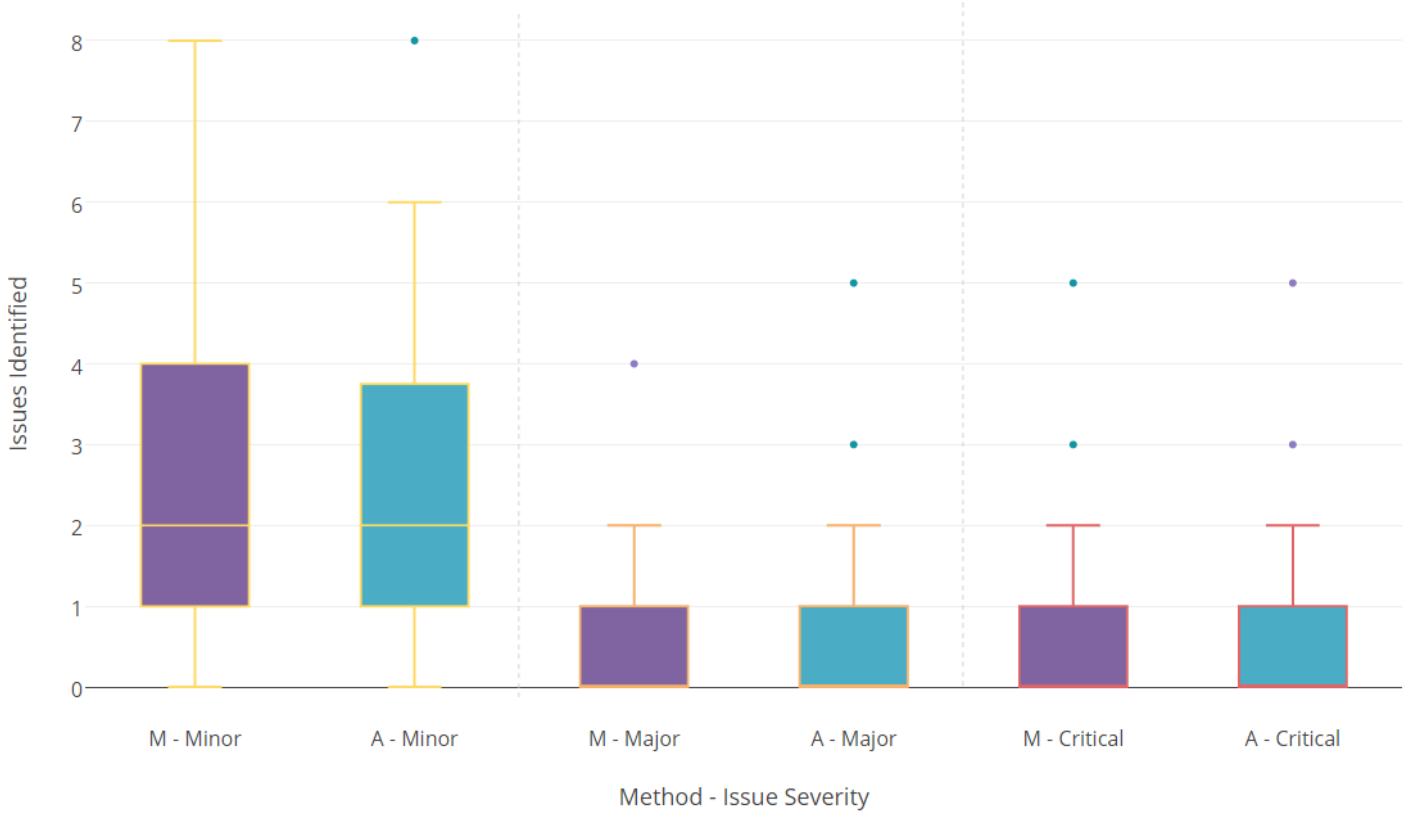
The findings indicate that participants in the *automated* research method were firmly convinced of what they had to say. Both groups are **equally open to self-reflect,** offer their **own ideas**, and ***critique*** the material **at near identical** rates. Similarly, presence of a moderator will not make participants more or less happy to be participating in the session. Bottom line: *Remote Automated participants can be just as invested in the study*. In *many ways* having a moderator will not create noticeable increases in participant feedback when the remote automated sessions are utilizing a strong Talk-out-loud oriented panel.

This data demonstrates the power and efficacy of UserZoom's new panel at delivering unprovoked, self-directed talk-along data. The top 3 reasons that full audio playback (TOL) is requested and desired for research are to aid in emotional deduction and proof-of-findings validation. Stakeholders prefer to see interactions from a *personal, empathetic* standpoint so they can inform future designs in relation to participant needs. 'Personal Moments 'along with emotional valence being equal, there is no doubt that remote automated participants are perfectly capable of feeling passionate towards the testing material and in conveying their answers in a similar consistency as moderated participants.

Remote automated research still explores Usability real well

As stated earlier, this research sought to explore a more holistic appreciation of feedback and not just focus on usability. As such, we only measured usability as 'issue awareness' on the part of the participants themselves. Given the presence of a moderator or lack thereof, the analysis focuses on how tuned the participants were themselves towards finding and mentioning usability issues. *In this regard, there was* ***no significant difference*** *across minor, major, or critical issues found.*

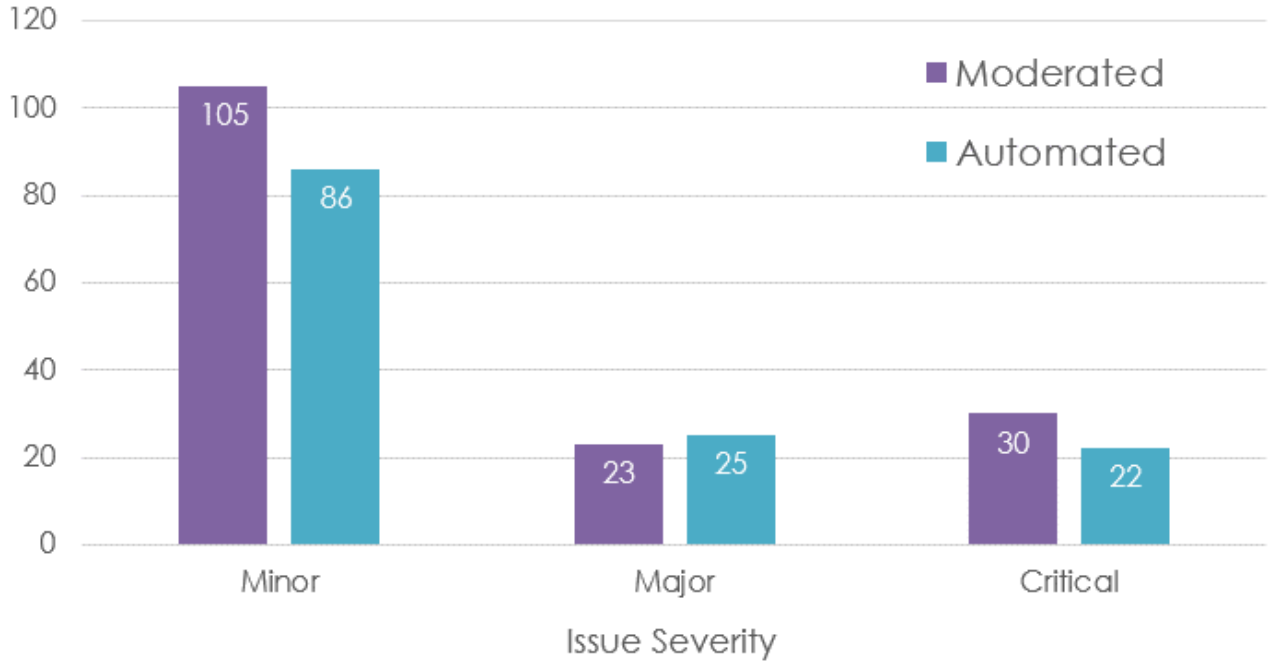
[Jump to the bottom to see how we defined Usability severity]



Despite some outliers among both methods, the 'typical' session found each participant self-identifying 3-4 issues. These were mostly 'Minor' ones according to our severity definitions.

Remember, this study did not look at how the two methods may have impacted expert inference or deduction of usability issues from the experiences. Instead it focused more on how the participants differed in their *own self-identification* of issues. These findings are in agreement with existing research looking at Issue Awareness and existing work on remote vs. moderated in the context of usability metrics (Task times and task completion rate). There were some really discrete differences in how frequently participants found certain issue types: there *were* slightly more minor issues found by moderated participants, but mostly due to several 'super-stars' that found 3-5 issues as opposed to the more common 2-3 minor issues found. Automated participants tended to be more subtle with where they went and what they said, so 50% ended up mentioning 1-2 usability issues. Major and Critical severity issues were near identical in their detection and mentioning.

Despite a handful of participants skewing the 'curve', the variances did not come out in favor of any method, for any issue type. The chart below shows the similarity of the total counts, as well as the previous box-plot above that demonstrates near identical distribution.

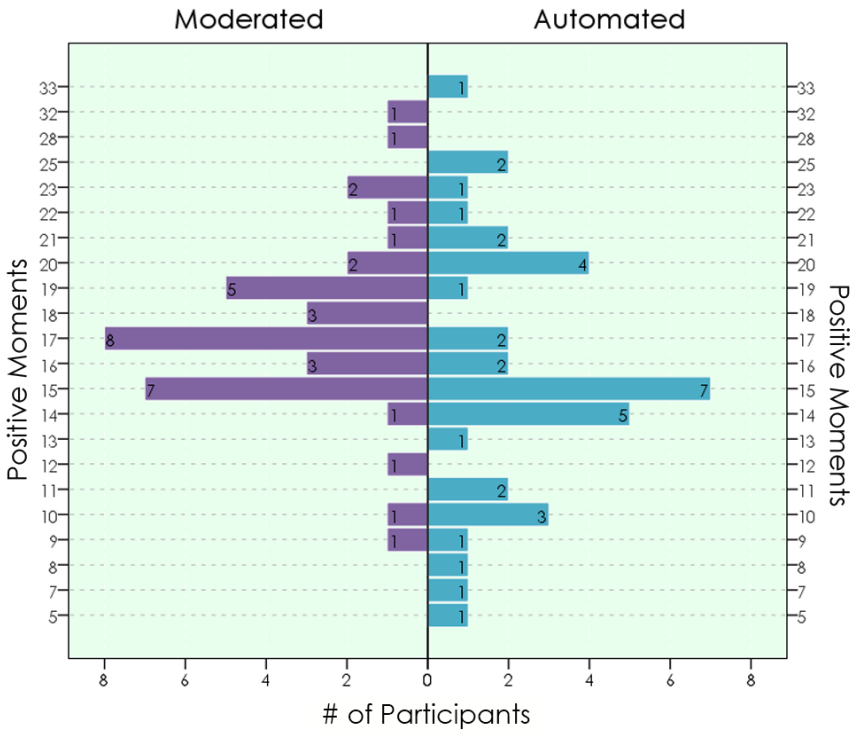


Includes repeatedly identified issues (Non-Unique issues counted)

There was a consistent emotional experience between the two methods

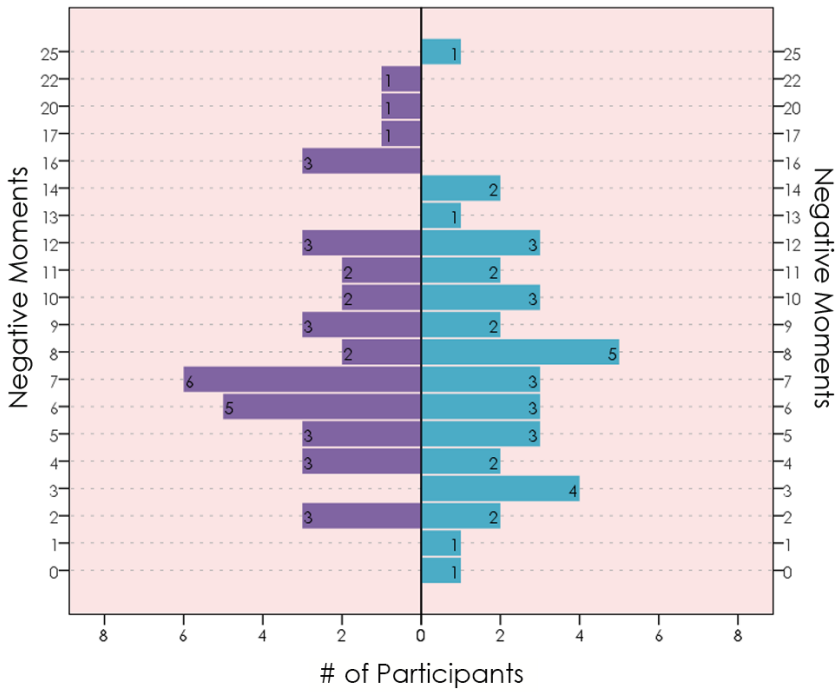
Sentiment strength and emotional variance are not only good predictors of salience (impactful feedback is often emotional feedback), but also of participant involvement and concentration. While my research did not go too deep in analyzing broader aspects of emotion (disgust, anger, etc.) looking at the high level variation of Positive and Negative moments was an acceptable starting point given the current lack of this available data.

For both Positive and Negative moments there was no difference between the conditions, indicating no effect on participant emotion based on moderator presence or lack thereof.



Despite high variance within each condition, emotional valence was actually one of the most consistent measures between the two methods (Avg. of 18/9 for Moderated, 16/8 for Automated).

These charts on the right depict the *rate* at which participants had a certain volume of positive or negative moments. For example, eight participants within the Moderated condition had 17 positive moments. The distributions are very similar across both methods, but what you can see is that for both positive and negative moments the variation is higher for the automated method.



Many benefits regarding remote unmoderated methods have been documented, but this may be the first glimpse into showcasing that participants 'feel freer to be themselves' without a moderator present. Tighter clumping on the moderated side suggests some participant assumption towards an expected response, however at this time the significance is non-existent in either direction. This will be discussed further below when the results from the 'Praise' moments are discussed.

Equal variances in 9/11 feedback 'moments' across both methods strongly implies that a quality study design (well-paced, open-ended, suggestive, etc.) and a targeted participant selection will impact feedback in these areas more-so than which tool and context you use.

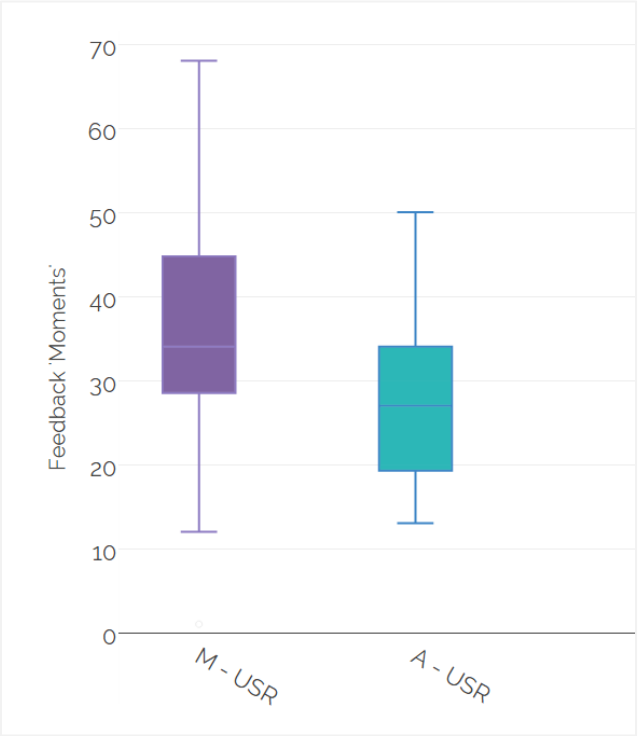
In other words, when the study context allows for a remote method making sure your sample size is adjusted properly and the participant pool is the best fit for your questions will impact overall data strength to a greater amount than whether a moderator is there to improvise some follow-up questions. While duration of session & amount of sentences may grow due to the moderator presence, its highly likely that these components do not lend added breadth or even depth to your data. ***But what will you gain?***

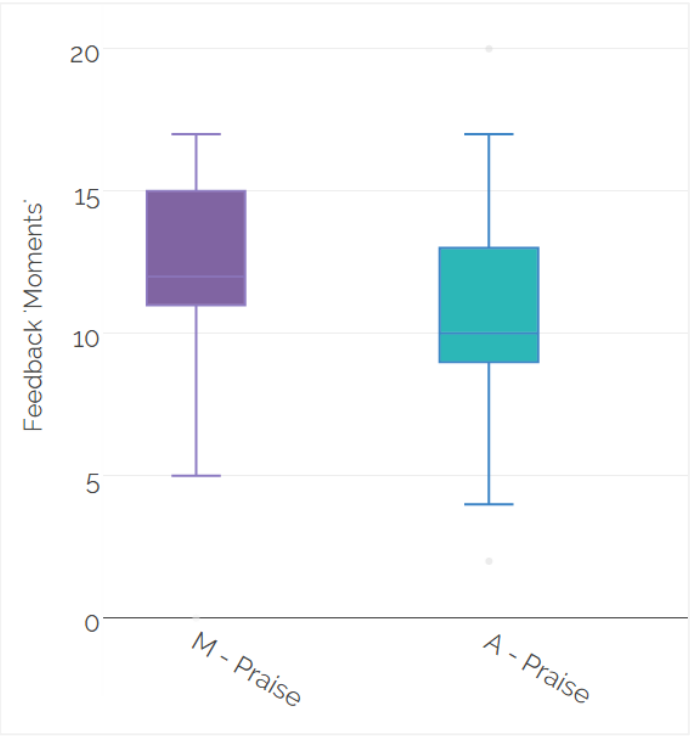
Two types of moments showed significant differences

The two methods resulted in significant differences between Unique supported reasoning and *Praise*.

While it's true that the moderated session resulted in a significant boost in the amount of explanation participants gave for their opinions (USR), we do need to be careful as researchers can place too high a priority on multiple explanations for the *same* initial point; while it might raise researcher confidence in the findings, it doesn't raise the total volume of findings (*conviction,* discussed a bit further down, was the metric used to track unique points that participants felt strongly about).

The significant difference in the *praise* feedback type implies support towards previous concerns on the presence of a moderator increasing acquiescence bias. It's highly likely that increased rate of *praise* during moderated settings is given to comfort the moderator or administrator. In our study, the presence of a moderator or authority lead to higher praise of the website and concepts.





Moderated participants explained their points further (due to moderator presence or follow-up probes), however the personalization throughout the sessions remained equal across methods.

It is highly likely that the clumping of emotional frequency displayed earlier relates to the tendency to appease a moderator with forced involvement. It's also likely that even the best moderators will fall into a habit of utilizing the same follow-up probes and discussion pacing. This is both a benefit (consistent, deeper probing & validating) and a curse of the human impact on the interviewing process.

As you'll see on the right with the additional avg. session stats, the moderated sessions *certainly* brought about more dialogue volume and participant feedback. Despite the moderated participant recordings lasting an average of 17 extra minutes, so much of the time during a moderated session is devoted to instructions, set-up, and clarifications that don't necessarily strengthen the data.

Avg. Session Stats	
Moderated	Automated
Avg | Median | Max	***Avg | Median | Max***
*Interview **Duration** (Min.)*	
39 | 34 | 56	**22 | 26 | 42**
Prompts & Tasks	
22 + ~8 follow-ups	**Scenario + 22**
*Response **Volume** (Sentences)*	
290 | 305 | 434	**228 | 241 | 373**
Age	
28 | 33 | 45	**32 | 35 | 42**
HH Income	
$55k | $65k | $140k	**$55k | $70k | $160k**
Muscial Relevance (1-7, Indifferent - Professional)	
5.4 | 5.8 | 7	**5.3 | 5.7 | 7**

At the same time the participant panel used for this study, while controlled to disqualify 'YouEye Pros' (we only took those that had done between 3-8 previous tests), is a participant panel well-versed in speaking to their computers rather than a moderator. In the next section, we'll discuss when you *should* run a moderated study, despite these findings demonstrating minimal data outcomes.

What Does it All Mean?

We're pretty excited about what this means for remote automated research, and clarifying options for researchers.

Make no mistake that moderated studies are important. And they also have their place. The decision between running a remote moderated study and a remote automated study should be based on considerations like:

- The contents of the study and if it's über confidential
- If the participants require more guidance, for example, an elderly panel
- If a concept is difficult and needs more explanation
- If the feedback is based on a physical product
- Budget, resources and time constraints

What we can say with certainty is that the decision *shouldn't* be based on fear of remote automated research producing less data, invalid data or uncontrollable data. This study helps to prove that. The biggest consideration: the quality of the study design. Writing a really well-crafted study for automated research takes understanding of the system being used.

There are considerations that need to be discussed when there is no moderator. And things like character limits, formatting and word usage come into play. Luckily, UserZoom has an FAQ section solely devoted to understanding this subtle crafting.

There is something else: Price considerations.
Different automated tools offer a range of solutions. Enterprise solutions for full executive reporting may cost the same as running an in-house moderated study where you can immediately jot down participant results as design memos or notes. This may be more efficient for your team than just signing onto a large proposal with an external vendor. However, if we're looking at the *capture component* alone, without assessing any bonus fees for analysis and reporting, than the automated method rules supreme when it comes to speed and cost.

Comparative Costs 40 Participant Study		MODERATED	AUTOMATED
Set-up *(Tool or Platform)*	*Cost*	$100-$2000	$49-$1000
	Time	*5-10 days*	*1-3 days*
Recruiting *(Panel access & result collection)*	*Cost*	$200-$1000	$0-$550
	Time	*5-16 days*	*2-3 days*
Incentives	*Cost*	$900-$1500	$600-$800
	Time	*1-3 days*	*N/A (automated)*
Moderator *(Hiring, ramping up, session facilitation)*	*Cost*	$1500-$2500	0
	Time	*5-16 days*	*0 Days*
TOTAL	***Cost***	**$2700-$7000**	**$800-$1800**
	Time	**5-16 days**	**2-3 days**

Given the variety of research solutions in the industry, a range of costs is displayed in the graph above.
Note: Just the capturing component of the study.

The cheapest moderated study will probably cost the same as the pricier automated solutions, when accounting for staff resources (recruiting, scheduling, tool testing, piloting, etc.)**.** Given the minimal impact of the data outcome, understanding when the moderated method *would* be the good choice for your research can help you save thousands of dollars.

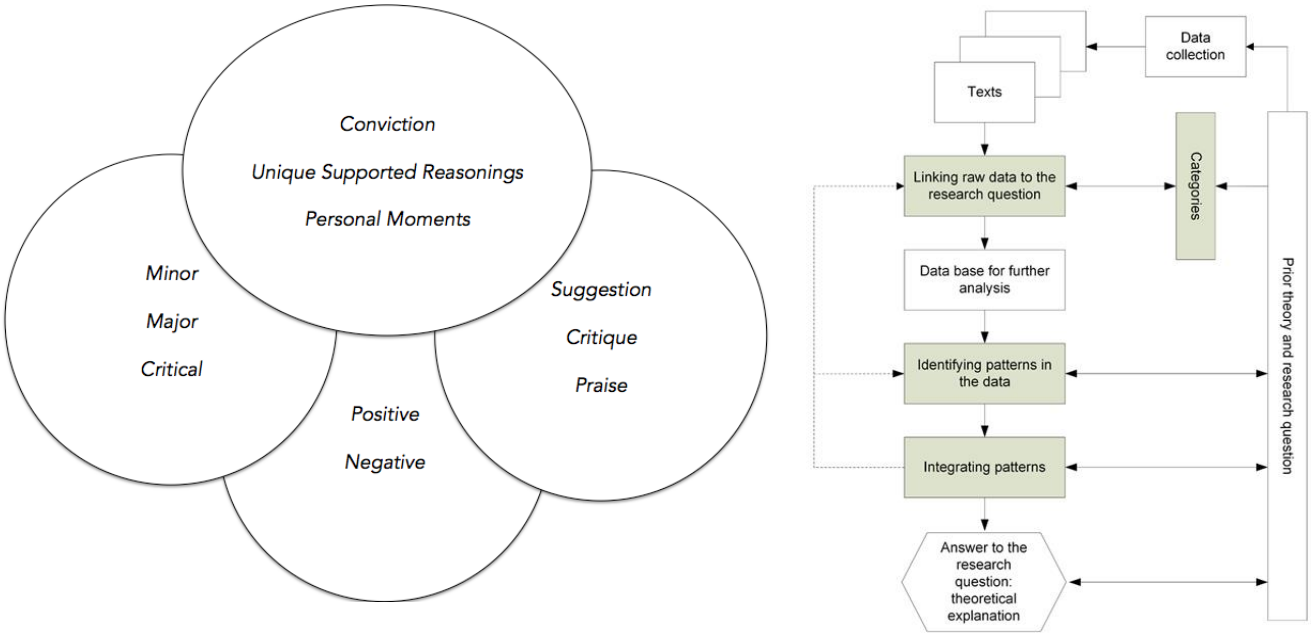
I'm aware that internal support (recruiting partners) and infrastructure (testing labs) may support ease of running moderated research, but as the tech world continues innovating, you should ask yourself: Can I innovate my market research as well?

For questions related to the research or a desire for more detailed findings & statistics, please contact
aackerman@userzoom.com

APPENDIX

About the coding framework…

Several in-depth sessions of transcript coding and interrater reliability were conducted to create the initial framework. After two weeks of developing a code set, we landed on 11 metrics to track and gauge these clusters; for instance the amount of *conviction* or *unique supported reasoning* (USR) participants held, and other areas, which we'll discuss later in this post.



Creating a coding framework for assessing feedback salience, relevance, and strength.

- **Conviction -** Four components: Strong words, strong emotion, repetition and multiple examples to explain a single point.
- **Unique supported reasoning (USR) -** Rationales, anecdotes or arguments used to explain an earlier opinion; how many separate unique points to support their answer for a given main point.
- **Personal Moments -** Unique (does not count redundant anecdotes) "events" within the feedback that relate to participant's own personal lifestyle, family or relationship experiences.
- **Positive emotional valence -** Four components: Good-natured words (praise), excited voice, smiling and laughter (related to prompt, not side banter).
- **Negative emotional valence -** Four components: Disgruntled attitude, sad voice, frowning and frustration.
- **Suggestion -** Any recommendation for how something should be different.
- **Critique -** Strong disapproval of material, topic or figure based on perceived faults or mistakes. Had to be directed towards some tangible thing rather than personal reflection.

- **Praise -** A compliment or admiration towards material, topics or figures (e.g., Designs, images, interactions or service).

- **Usability Issues Encountered**
    - Critical - A complete block of goals and struggle to understand what's needed by system. If a solution never found, definitely Critical. Even if found, 'Critical' if participant had to attempt 3+ other paths or delayed beyond 1 minute.

    - Major - A loss of function that completely stops workflow and does not provide an easily known workaround. Workaround found, but with slight delay.

    - Minor - Cosmetic or general annoyance that does not impact a goal, but may delay, however workaround found quickly.

Significance score for all 11 feedback metrics….

Defining Remote Moderated and Remote Unmoderated

If you're already familiar with remote moderated research and remote unmoderated research, you can skip down to the next section for the findings. However, understanding the variables intrinsic to the two methods can strengthen our understanding of what would cause differing feedback outcomes. Keep in mind that when I reference 'remote unmoderated' research I am implying the usage of the talk-out-loud protocol, not just the recording of passive backend data collection.

Remote *moderated* research

A favorite amongst companies over the past 20 years or so that want to reduce the overhead associated with other qualitative research methods like focus groups. An online conferencing tool connects the moderator (typically drives the pace of the interviews) with the participant. The moderator introduces what is about to happen, speaks the prompts aloud, can view the participants' screen and their faces via webcam, and is allowed to improvise follow-up questions and clarifications.

For the moderated condition in this study, the moderator could improvise and probe deeper as long as they…

a) Did not introduce novel concepts and

b) New probes did not lengthen the session beyond the 1-hour allotted per session.

The tool used for these moderated sessions was Citrix's GoToMeeting.

Remote *unmoderated* research

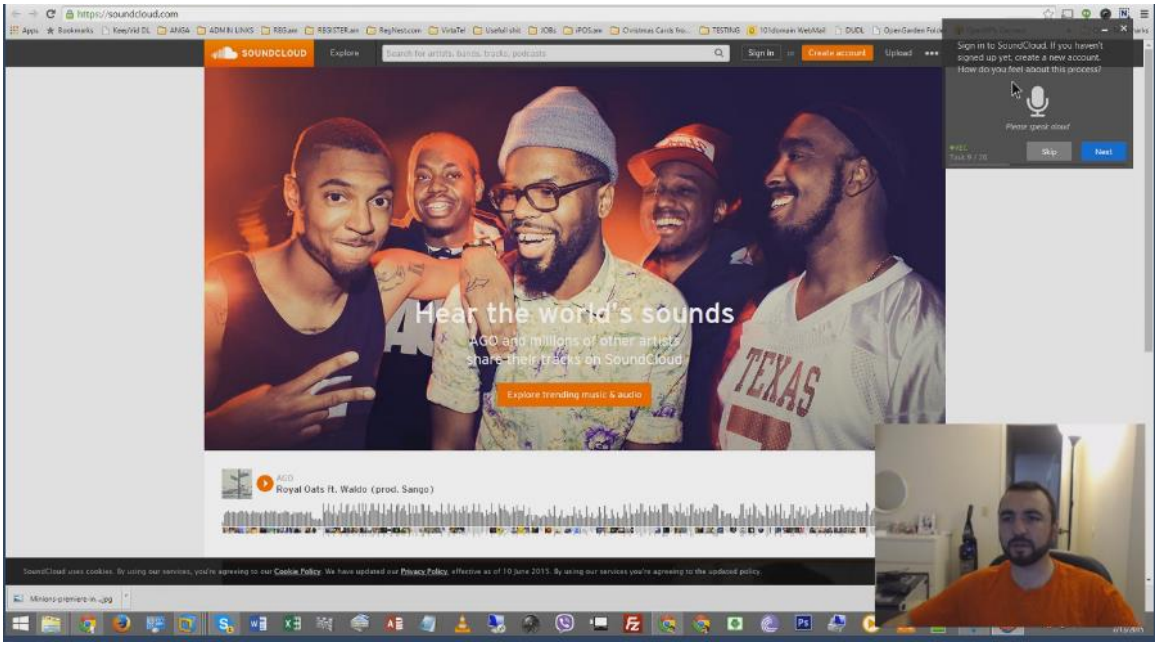
Newer to the scene and gaining ground, this type of research allows participants to guide themselves through a research study remotely using video capture technology and their device (mobile or desktop) through pre-defined prompts. The pre-study scenarios are typically brief and participants can see how many tasks are left at any given time (e.g., x/25). No set time is given to complete the session. The automated software sends participants to a pre-determined website, and the tasks may also ask for single choice, ranking, Likert ratings, or other closed-ended input desired.

Participant experience…

Running some deeper analysis displayed a *slight* trend towards moderated participants feeling like they were making a larger contribution, though still at insignificant differences. One issue with this measure

is that despite the moderator not recording the session, the moderator was still present and watching via screen share as they input ratings.

Moderator presence may have led to different ratings given the “meta” nature of these self-reported questions. All self-reported measures were taken as a seven-point Likert rating. There was a significant difference in the Ease of Concentration between the two methods, with moderated participants finding it easier to focus on the tasks at hand.



Footnotes (Academic research articles)

1. Brush, A.J. Bernheim; Ames, Morgan; Davis, Janet. 2004. A comparison of synchronous remote and local usability studies for an expert interface, CHI '04 Extended Abstracts on Human Factors in Computing Systems, April 24-29, 2004, Vienna, Austria

2. Brunn, Anders. Stage, Jan. 2012. The Effect of Task Assignments and Instruction Types on Remote Asynchronous Usability Testing. Session: Usability Methods CHI 2012, May 5–10, 2012, Austin, Texas, USA

3. Rob Martin, Majed Al Shamari, Mohamed E. Seliaman, and Pam Mayhew. 2014. Remote Asynchronous Testing: A Cost-Effective Alternative for Website Usability Evaluation. International Journal of Computer and Information Technology (ISSN: 2279 – 0764) Volume 03 – Issue 01, January 2014