Working paper, draft for submission (HCI / CSCW family) · June 2026

The Lookup That Assumes Competence: A Credibility-Gated, AI-Moderated Study of How Visitors Use a Theme-Park Mobile App

Nicole Zeng,[1] Benjamin Lo,[2] and Justin Hoang[2]

[1] *The Observatory* [2] *Dialogue AI*

Corresponding author: Nicole Zeng, gnezelocin@gmail.com

Abstract

Mobile applications increasingly mediate complex physical experiences, yet we know little about how their information architecture meets, or fails, the moment-to-moment demands of a crowded, time-pressured day. We report a qualitative study of 29 recent visitors to a large theme-park resort who used its official mobile app before and during their visit, conducted with an AI moderator and analyzed with an ethnographic, priors-quarantined coding protocol. We make two contributions. Empirically, we describe a recurring structure we call the competence assumption: the app functions as a real-time lookup that performs well for users who already hold a working model of the park and offers little to those still building one, so that strategy migrates to external sources (Reddit, social media, staff) and the deficit among newcomers is latent, unfelt by the newcomer and witnessed instead by experienced companions. Methodologically, we contribute a per-interview quality apparatus for AI-moderated research, the Interview Quality Score (IQS), which decomposes data quality into participant signal, moderator effectiveness, and data usability, gates signal depth on authenticity, and pairs with a verified-verbatim provenance chain so that every quote is located against source. We position both contributions honestly: this is a small, single-market, cross-sectional study, the findings are descriptive rather than causal or population-general, and the prevalence of qualitative patterns is reported as a presence floor rather than an incidence rate.

Keywords: AI-moderated interviews, qualitative methods, information architecture, human-computer interaction, data quality, theme parks, mobile UX

1. Introduction

People increasingly run consequential parts of their lives through an app while standing inside the situation the app describes. A theme-park visit is a sharp instance: the day is time-pressured, crowd-dependent, socially coordinated, emotionally loaded, and constantly changing, and the official app is meant to help the visitor navigate, plan,

eat, enter, and queue. The design question is not only whether a visitor can find a feature, but whether the structure of the app matches how a park day is actually lived.

Most evaluation of such apps relies either on instrumented usage logs, which record what was tapped but not what was meant, or on satisfaction surveys, which record a rating but not the contradiction beneath it. Both miss the texture where the design succeeds or fails: the moment a parent stops to read a screen while a toddler pulls at them, the moment a first-time visitor cannot find the entry pass at the gate, the moment an experienced user defends a cluttered home screen they have simply learned to tolerate.

Contribution statement. This paper contributes two things at the level the evidence supports. First, a descriptive empirical finding: the app operates as a competence-assuming lookup, and the resulting orientation gap is latent in newcomers and observable mainly through the accounts of experienced companions. Second, a methodological apparatus for trusting AI-moderated qualitative data: a per-interview Interview Quality Score with an authenticity gate, coupled to a verified-verbatim provenance chain. We claim neither causation nor generalization beyond this sample.

The remainder positions the work against prior research on AI-moderated interviewing and the limits of self-report, describes the AI-moderation and coding protocol, presents the findings organized by structure rather than by interview, and closes with the threats to validity that bound the claims.

2. Related work and background

A fast-growing literature uses large language models to conduct or support qualitative interviews. De Paoli (2024) demonstrates and probes the limits of inductive thematic analysis performed with an LLM; Jarrahi (2025) frames the broader turn toward conversational AI in qualitative inquiry; recent systems work develops adaptive LLM interviewers and compares them to human-led interviews (*AI Conversational Interviewing*, arXiv:2410.01824, 2024). This work establishes feasibility and surfaces a central worry: when the interviewer is a machine, and when generated text can imitate a participant, how do we know the signal is real? Studies examining LLMs as simulated participants (*Simulacrum of Stories*, arXiv:2409.19430, 2024) make the authenticity problem concrete.

That worry has an older root. Self-report has never been a transparent window onto behavior. LaPiere (1934) documented the gap between stated attitude and enacted action; Goffman (1959) separated the front-stage performance from the back-stage conduct; Loftus (1975) showed how a leading question reshapes a recollection. Spradley (1979) argued that meaning lives in the participant's own term, not the researcher's imported category. AI moderation inherits all of these hazards and adds one: a fluent machine can both prompt agreement and generate plausible detail.

What the literature lacks is a standard, per-interview way to express how much to trust a given AI-moderated transcript, decomposed so that a low score tells you *where*

the failure lies (the participant, the moderator, or the data), and tied to a provenance chain that lets a later reader verify each quoted line. On the substantive side, the design literature has long known that systems can be feature-rich yet hard to discover (Norman, 2013), and that behavior is produced by the surrounding structure rather than the individual's character (Meadows, 2008). We extend both into the specific setting of an experience-mediating app used under time pressure.

3. Methods

Design. Twenty-nine recent visitors completed a single AI-moderated conversational interview (roughly 10 to 18 minutes) on the Dialogue AI platform. The moderator followed a discussion guide that anchored on a real remembered moment ("think of one time you opened the app, what was happening"), then probed feature expectation against two screen stimuli (the home and tab structure, and the home broken into labeled sections), information gaps, the new-user learning gap, and a forced single-priority close. A documented moderation principle instructed the moderator to avoid leading toward any "guided mode" answer and to use neutral probes; we treat the resulting unprompted appearance of guidance requests as evidence rather than artifact (see Results).

Sample and recruitment. Participants were screened for a resort visit within roughly 48 hours, entry to at least one park, and personal app use. The achieved sample (Table 1) is 5 first-time and 24 repeat visitors across five frequency segments; the study planned 60 and achieved 29, a limitation we treat as load-bearing. The sample skews California-resident (21 of 29) and iPhone (23 of 29).

Table 1. Sample and segments (N = 29)

Segment	n	Note
First-time	5	directional only; ~3 clean (P03 wrong-park flag, P05 truncated). Often reported ease (carried by companion or surprise-preferring).
Repeat-Occasional	7	'special occasions'; mixed reliance; includes lapsed returners (P09, P28) who behave like quasi-first-timers.
Repeat-Regular	2	~once a year; too few to characterize.
Repeat-Frequent	9	multiple times a year; optimizers and food-discovery critics concentrate here.
Repeat-MagicKey	5	passholders; span satisfied to frustrated; LL/VQ confusion reaches this tier.
Repeat-Unspecified	1	P29, screener metadata missing.

Single-method qualitative; no quant instrument. First-time cell is directional (see Limitations).

Analysis. Each transcript was read whole, then coded bottom-up in participant language (open to axial to synthesis), blind to a priors ledger that was extracted from

the study plan before coding and scored only afterward. This quarantine is the study's structural defense against findings that merely echo the brief. The analysis engine takes a position; a downstream normalization step then flattens that position into a neutral, audience-free findings package carrying every finding with its disconfirming evidence alongside, which is the artifact this paper renders.

Quality and credibility. We scored each interview on the Interview Quality Score (IQS): Participant Signal Depth (weight 0.50), Moderator Effectiveness (0.30), and Data Usability (0.20), each on a 0 to 5 scale, composited linearly. A rule sits above the components: signal depth is gated by authenticity, so that an interview flagged for probable AI assistance or confirmed out-of-market caps its signal at 3 regardless of apparent richness. Table 2 reports the study-level summary. The full per-interview table (P##-anonymized) is carried in the data bundle and rendered as a credibility waffle in Figure 1.

Table 2. IQS summary (weights, component means, distribution)

Component	Weight	Mean (of 5)	Note
Participant Signal Depth	0.50	3.86	gated by authenticity (Usability=1 caps Signal at 3)
Moderator Effectiveness	0.30	3.76	softest channel: occasional forced-choice probes
Data Usability	0.20	3.9	carries the authenticity flags
Composite	,	**3.84**	range 2.3–5.0; bands: 7 high / 13 solid / 8 caution / 1 low

Composite mean derived from the 29 per-interview rows, not stated independently. Authenticity cap bound one interview (P03, out-of-market conflation), whose signal was already 3.

Figure 1. IQS credibility waffle, one cell per coded interview (N = 29)

Color encodes composite band; dashed outline marks an authenticity or completeness flag. The three flagged-and-cautioned cells (out-of-market conflation, truncated transcript, context inconsistency) all fall in the first-time cell, so the thinnest cell is also the least clean.

4. Results

We organize by structure, not by interview. Because N = 29 sits below a conventional threshold for percentages, we report no incidence rates; where we give a count it is a presence floor (the number of participants who raised something unprompted), not a measured rate, and most items were not asked of all participants.

4.1 The competence assumption. Across experience levels, participants describe the app as a fast lookup for wait times, the map, food, and tickets that works for someone who already holds a model of the park, and is thin for someone still building one. One frequent visitor, defending the cluttered home screen, drew the distinction precisely: "everything has its purpose [...] I personally wouldn't move anything. But [...] if it was for someone that didn't have much experience, maybe they would want these things on the home screen" (P15, located in the bank). A first-time visitor named the same structure from the other side: the app was "for someone that already knew how to like figure out the app." Satisfaction among veterans is therefore better read as survivorship than as evidence the design teaches.

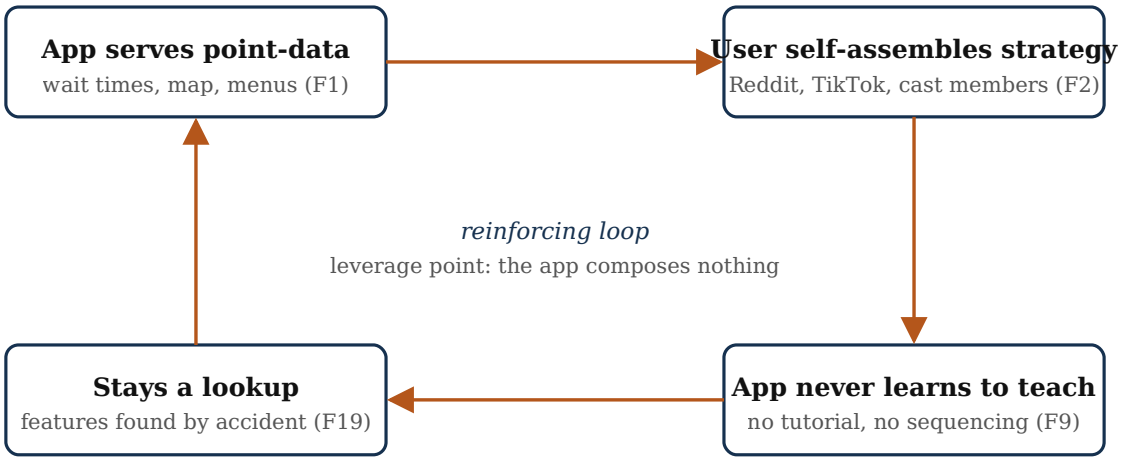
4.2 Strategy migrates outside the app. Participants report learning park strategy, Lightning Lane timing, ride sequencing, insider rituals, from Reddit, social media, cast members, or a knowledgeable companion, and do not report the app teaching it. One participant stated the absence directly: the app "doesn't actually really teach you that virtual queues exist [...] there's no tutorial." The requests for a guided mode, tutorial, route, or itinerary recurred and arose without the moderator prompting them, which (given the documented anti-leading principle) we read as a participant-originated desire rather than a moderation artifact.

4.3 The deficit is latent and witnessed. The most detailed accounts of first-time difficulty came not from first-time participants, several of whom reported the day as easy, but from repeat visitors describing companions: friends who "scramble to scan their phone because they don't know where it is" at the gate, a cousin who found the digital ticket only by accident. We state this as a mechanism rather than a rate: the orientation gap is unfelt by the person who has it and observable through the experienced person beside them. A disconfirming case is carried alongside (one repeat visitor reported his three first-time companions had no trouble, "we're all pretty app savvy"), bounding the claim: newcomer ease is real for some and tracks prior digital fluency.

4.4 The queue system as a comprehension gap. Lightning Lane, Virtual Queue, and reservation mechanics were described as unclear across experience levels, including by annual-pass holders, one of whom could not say what Virtual Queue is. Because access to these features is a paid mechanic, the comprehension gap has a second-order implication a usability frame misses: confusion plausibly suppresses uptake of a revenue feature, a connection one passholder drew unprompted.

Figure 2 renders the relationship among these findings as a single reinforcing structure.

Figure 2. The competence-assumption loop



A reinforcing loop (Meadows, 2008): point-data without composition pushes strategy outside the app, so the app is never pressured to teach, which keeps it a lookup. The low-order leverage point is that the app composes nothing.

4.5 Concrete frictions. A support tier of specific, recurring frictions accompanies the structural finding: map markers that require zoom and tap to identify and do not orient to facing; restrooms as the most frequently named concrete navigation gap (a presence floor of roughly 8 of 29), with no easy filter; food discovery that proceeds vendor-by-vendor with no cross-park food-first view; the entry pass and gate scan not surfaced prominently; and wait-time inaccuracy or lag that breaks the moment. These are reported as a floor, not a rate.

4.6 A presence cost. Participants described app use during the visit as simultaneously necessary and a cost, captured in one account as "exhausting and necessary." Making time legible (live waits, booking windows) appears to convert not-checking into felt waste, and the participants most troubled by phone use were sometimes the heaviest users. We carry the disconfirmation: a subset rationed app use by choice to stay present, matching word to deed.

5. Discussion

The competence assumption reframes a common design instinct. The intuitive response to a confused newcomer is to add features or surface more content; our data suggest the binding constraint is the opposite, the app already holds the data a visitor needs but composes none of it into guidance, so adding data fields does not address the deficit. The leverage point (Meadows, 2008) is the composition rule, not the feature

count: a sequencing or teaching layer changes the system's behavior where another data field does not.

The latent-and-witnessed structure of the deficit has a methodological consequence beyond this study. If newcomers under-report their own difficulty, then a survey of newcomers will understate the orientation problem, and the more reliable signal comes from the experienced people who watch them struggle. Designing the new-user experience from veteran observation rather than newcomer self-report is a practical implication that follows directly.

For AI-moderated research, the apparatus we report addresses the authenticity worry the method literature raises. A per-interview score that decomposes quality and gates signal on authenticity lets a study say not just "how good was the data" but "where did it fail and how much should this transcript anchor a claim," and the verified-verbatim chain makes every quoted line checkable against source. We offer IQS as a candidate instrument for that purpose, not as a validated scale; establishing its inter-rater reliability and convergent validity is future work.

6. Limitations

The threats to validity are real and bound every claim above. **Sample and coverage:** the achieved N is 29 against a planned 60, and the first-time cell is 5, of which roughly 3 are clean (one is an out-of-market conflation excluded from primary first-time evidence, one is truncated). First-time conclusions are directional and lean on veteran observation. **Cross-sectional:** a single interview per participant supports no causal claim. **Self-report and say-do:** much of the data is recalled and stated, with the documented hazards (LaPiere, 1934; Goffman, 1959; Loftus, 1975); we mitigate by anchoring on specific moments and by carrying behavioral evidence where available, but the gap between stated and enacted remains. **AI-moderation effects:** the moderator occasionally used forced-choice probes that can steer, and two screen stimuli failed to display at full fidelity (one too small to read, one not matching a logged-in screen), which means the home-screen clutter finding is stimulus-anchored and likely overstated relative to a personalized real screen. **Single market:** a California-skewed, iPhone-skewed, single-resort sample does not generalize to other parks, platforms, or regions. **Prevalence:** qualitative counts here are presence floors, not incidence. None of these is fatal to a descriptive contribution; all of them forbid a causal or general one.

7. Conclusion

A theme-park app that answers "what is the wait, where is this, what is on the menu" can still fail the visitor who does not yet know how the day works, and that failure is hardest to see precisely because the people who have it cannot feel it. We have described that structure at the level our small, single-market, cross-sectional data

supports, and we have offered a per-interview credibility apparatus for trusting AI-moderated qualitative data and a provenance chain for verifying its quotes. Both are starting points, the empirical one for design that teaches rather than merely informs, the methodological one for a field that will increasingly let machines do the asking.

References

De Paoli, S. (2024). Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. *Social Science Computer Review, 42*(4), 997–1019. https://doi.org/10.1177/08944393231220483

Goffman, E. (1959). *The presentation of self in everyday life*. Anchor Books.

Jarrahi, M. H. (2025). Interviewing AI: Using qualitative methods to explore and capture machines' characteristics and behaviors. *Big Data & Society.* https://doi.org/10.1177/20539517251381697

LaPiere, R. T. (1934). Attitudes vs. actions. *Social Forces, 13*(2), 230–237.

Loftus, E. F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology, 7*(4), 560–572.

Meadows, D. H. (2008). *Thinking in systems: A primer.* Chelsea Green.

Norman, D. A. (2013). *The design of everyday things* (Rev. ed.). Basic Books.

Spradley, J. P. (1979). *The ethnographic interview.* Holt, Rinehart and Winston.

AI Conversational Interviewing: Transforming surveys with LLMs as adaptive interviewers. (2024). arXiv:2410.01824. [verify authors/venue before submission]

Simulacrum of stories: Examining large language models as qualitative research participants. (2024). arXiv:2409.19430. [verify authors/venue before submission]

[CITATION NEEDED: a theme-park or queue-management HCI reference to position the applied setting, if the target venue expects domain grounding.]

Appendix A. Interview protocol (probe areas)

Summary of probe intent, not the verbatim script and not participant answers. (1) Warm-up: visit context (who, which parks, what kind of day). (2) Anchor on one real app-opening moment; probe what was opened first, why, what was expected, whether it worked, what came next. (3) Frequency and deliberate non-use. (4) Tab expectation against the home/structure stimulus. (5) Information gaps: hardest to find, easiest to miss, checked repeatedly. (6) Home deep-dive against the labeled-sections stimulus: useful vs. promotional, move or remove. (7) New-user learning gap. (8) Forced single-priority close. Moderation principle: neutral probes, avoid leading toward a guided-mode answer.

Appendix B. Codebook summary (synthesis codes)

Table B1. Synthesis codes (qualitative; not sized)

Code	Label	Definition	Prevalence
S1	The competence assumption	app assumes you already know how to run a park day	not sized
S2	Generic vs. mine	home reads as a billboard; users want it to be their day	not sized
S3	Phone-tether paradox	app is necessary and a cost to presence	not sized
S4	Data, not guidance	serves point-data, composes no plan	not sized
S5	Map shows, doesn't read	legible data, hard to read in the moment	not sized
S6	The queue tangle	Lightning Lane / Virtual Queue widely not understood	not sized
S7	Designated app operator	one person carries load and social cost	not sized
S8	Food discovery is a database	location-first, vendor-by-vendor	not sized
S9	Veterans witness the novice gap	latent deficit, observed by companions	not sized

Prevalence marked "not sized": at N = 29 the package reports presence floors, not incidence.

Appendix C. Representative verbatims (curated selection)

A confidentiality-safe selection, not the full bank. Every quote is exact and located (P## + demographics + timestamp). Full transcripts are withheld to protect participant confidentiality; this is a privacy measure, not a verification gap, the quotes below remain verified against source.

Table C1. Selected verbatims

Participant	Locator	Quote
P06 35-44 F CA	[04:02]	Both. It felt exhausting and necessary. I hated how much I had to be on my phone to make things work. I really despised that element of it.
P04 18-24 M CA	[15:05]	the app was like for someone that already knew how to like figure out the app […] a guided mode would have made everything better for me.
P12 18-24 M CA	[12:15]	the app doesn't actually really teach you that virtual queues exist […] there's no tutorial like, 'Oh, it's your first time at Disneyland, you should know this.'
P26 25-34 F CA	[04:34]	I don't know the things that I don't know [...] there's a lot of different tabs and clickable things on the app that I've never clicked.
P16 18-24 F CA	[12:34]	I've personally seen my friends […] literally wait for five minutes in front of the entrance trying to scramble to scan their phone because they don't know where it is.
P23 25-34 F CA	[02:54]	although I'm a big Disney fan […] I'm not super knowledgeable when it comes to the technicalities […] once you ride your first ride you can then book another ride. I don't know how that works.
P27 25-34 F CA	[15:52]	I kind of almost just beg […] put in something where I can see overall all the food options across the parks um without having to click into each vendor because that's so tedious.
P24 35-44 M NJ	[08:58]	No, we're all pretty app savvy.

Appendix D. Demographics

Table D1. Sample composition (N = 29)

Dimension	Distribution
Gender	Female 18, Male 11
Age	18–24: 11; 25–34: 8; 35–44: 6; 45–54: 4
Geography	California 21; out-of-state 8 (WV, TX, IL, GA, NV×2, NJ, OH)
Phone OS	iPhone 23, Android 4, Both 1, Unknown 1
Segments	First-time 5; Occasional 7; Regular 2; Frequent 9; Magic Key 5; Unspecified 1