Journal of User Experience

Vol. 20, Issue 2, February 2025 pp. 52–64

Who's There? Mystery Calling for UX Research

Michael J. Madson
Assistant Professor
Arizona State University
Santa Catalina Hall, 251U
7271 E Sonoran Arroyo Mall
Mesa, Arizona
USA
michael.madson@asu.edu

Yoshita Gade
User experience researcher and product designer
Arizona State University
6049 S Backus Mall
Mesa, AZ
USA
ygade@asu.edu

Unnati Srivastava
User experience researcher and product designer
Arizona State University
6049 S Backus Mall
Mesa, AZ
USA
usrivas4@asu.edu

Abstract

This article provides a critical discussion of how mystery calling can support user experience research, offering insights grounded in the extant literature as well as our experience as mystery callers. The pros of mystery calling include data from (near) authentic interactions, scalability, cost efficiency, and broad applicability. The cons include medium bias, rigidity, challenges with standardization, and possible consequences when participants learn of the minor deception. Yet, the telephone is not a stable technology, and we anticipate significant impacts from artificial intelligence (AI). Broadly, we anticipate that AI will make future mystery caller studies more dependent on machine-machine interactions, with the human playing more of a supervisory and editing role. In particular, AI will automate and expand functions related to data collection, data analysis, project management, and research reporting. Given the current state of the method and its trajectory, we conclude with practical tips for UX researchers who may deploy mystery calling in their own work. We emphasize the need to weigh the ethics of mystery calling carefully.

Keywords

Mystery calling, UX research, telephone, customer journey, artificial intelligence (AI), speech synthesis, text-to-speech (TTS), voice conversion (VC), natural language processing (NLP), ethics

Introduction

No user experience project is without challenges. Researchers may struggle to collect adequate data or generate insights from a limited sample (see Robinson & Lanius, 2018). They may need to balance multiple tasks with competing objectives (Da Silva et al., 2012; Hinderks et al., 2022). Especially in market downturns, they may need to navigate tight budgets, fast timelines, and questions over the value of their work. As a result, it is important to identify UX research methods that are low-cost and high-yield.

One of these methods may be mystery calling, a form of mystery shopping. Mystery shopping is a methodological umbrella for evaluating "any type of customer service process by acting as actual or potential customers (who) in some way report back their experiences, in a detailed and, as far as possible, objective way" (Turner, 2012, p. 333). Since the term was coined in the 1940s, mystery shopping has expanded conceptually, and under current usage, it may not involve actual shopping (Turner, 2012). It may involve, for instance, an overall assessment of a service's user friendliness and informational accuracy.

A distinguishing characteristic of mystery calling is conversations over a telephone, which provides perspective into business as usual. Mystery calling is adaptable across populations, places, and projects, and it is becoming common in sensitive settings like healthcare and finance. Yet, like other forms of mystery shopping, mystery calling raises ethical considerations because it involves minor deception, as Dickson et al. (2018) noted. Complicating matters, there are few accessible, peer-reviewed resources on mystery calling for UX research. In fact, few articles in the *Journal of User Experience* and its predecessor, the *Journal of Usability Studies*, have discussed the use of telephones (for example, Khan et al., 2016; Lamm & Wolff, 2021; Lewis & Sauro, 2021; Gardner-Bonneau, 2010); this represents a significant gap in our literature.

Here, then, we offer a critical discussion for UX researchers. First, we describe the pros and cons of mystery calling as a UX research method, capturing the state of the art. Second, we look ahead to how mystery calling may transform with the expansion of artificial intelligence (AI). Finally, we propose some best practices. These sections draw on scholarship across disciplines as well as our own experiences as mystery callers and UX professionals. Note that we use the term "participants" to indicate those who provide data in a study, knowingly or not. In particular, the participants in a mystery caller study are those who answer the telephone during a mystery call, usually company employees.

Pros of Mystery Calling

Mystery calling has at least four main advantages in UX research: data from (near) authentic interactions, scalability, cost-effectiveness, and wide applicability.

Data from (Near) Authentic Interactions

UX research methods are complicated by social desirability bias and the Hawthorne effect, which can negatively impact data quality and, in turn, company decision-making. Social desirability bias is a mismatch between what participants really think and what they present to the research team. It occurs when participants want to "look good," particularly when they address questions involving a controversial topic or perform tasks with clear social norms (Bergen & Labonté, 2020). A related phenomenon, the Hawthorne effect, occurs when participants know that they are part of a study and consequently modify their behavior. Either phenomenon undermines confidence in the findings. As a corrective, mystery calling catches participants by surprise, and the study remains a secret during the data collection. As a result, participants will probably interact with the researchers as they would with ordinary callers, increasing the authenticity of the data. The data can be quantitative, qualitative, or both.

Depending on the study, mystery callers can be members of the target population rather than full-time UX researchers. Van Hoof et al. (2014), for instance, investigated how easy it was for teenagers to purchase alcohol through home delivery services to assess compliance. They recruited 12 teenagers, 6 males and 6 females, and trained them in mystery calling. Then, data collection began. Surprisingly, the team found a dismal compliance rate with local laws. None of the home delivery services asked the mystery callers for identification documents, and only four complied with the Dutch Licensing and Catering Act over the phone.

With these qualities, mystery calling can help UX researchers map customer journeys, track adherence to standard operating procedures (such as at call centers), and conduct competitive analyses. Based on the findings, managers can set benchmarks and pinpoint areas for improvement (see Voxco 2024a, 2024b).

Scalability

UX projects can vary dramatically in terms of scope and data volume, which can complicate resource allocation. Mystery caller studies are highly scalable because they can support small or large project scopes, with modest or massive volumes of data. Consider a few recent examples. On the smaller end, Van Hoof et al. (2014) sampled 30 home delivery services for alcohol, as mentioned above. On the larger end, Gravlee et al. (2023) assembled a team of 22 mystery callers to contact 591 community pharmacies in Mississippi, United States. Bucher et al. (2023), a team of five researchers, made 1,383 mystery calls to German citizens. The scalability of mystery calling can stretch further with AI tools, as discussed below.

As a project scales up, mystery callers need not play themselves. They can adopt personas to simulate multiple customer journeys, as Corbisiero et al. (2023) demonstrated. Corbisiero's team members constructed four "scripted clinical vignettes" that could capture a range of contexts. The vignettes varied by subspeciality in obstetrics and gynecology (pelvic medicine and reconstructive surgery, cancer care, maternal-fetal medicine, reproductive endocrinology, or infertility), medical complaint (stress urinary incontinence, new-onset pelvic mass, kidney transplant, or infertility), age (35 years old, or 65 years old), referral source (primary care or emergency department), and symptoms. Taken together, the data from the vignettes enabled stronger conclusions than a single vignette would have on its own (Corbisiero et al., 2023).

Cost Effectiveness

Many companies have limited resources for UX research. Yet, mystery calling generally does not require participant remuneration, convenience fees, or room booking fees. Neither does it require specialized technologies such as user testing labs or analytics software. All it requires is a phone and a notetaking device, which can be as simple as paper and a pencil. In return, the findings can help companies define weak links, increase customer satisfaction and loyalty, and avoid errors, even legal sanctions in industries like banking (Kurtovic & Hasimbegovic, 2015). The results may also inform staff training.

Analysis of mystery call data tends to be straightforward. Quantitatively, researchers have captured easy-to-count metrics that include call duration, call attempts, and number of holds (Lungfiel et al., 2023), compliance rates (van Hoof et al., 2014), time to the next available appointment (Pollack et al., 2016), the immediate availability of certain products (Ditmars et al., 2019; Egan et al., 2019; Lungfiel et al., 2023), and task completion time (Giacomelli & Tonello, 2015), also known as "one of the most important metrics in usability tests" (Rummel, 2014). Qualitatively, researchers have identified types of user problems (Bucher et al., 2023), staff product recommendations (Dickson et al., 2018), and strategies and justifications for ethnic discrimination (Verstraete & Verhaeghe, 2020). Researchers have also evaluated conversation understandability (Bucher et al., 2023), as well as staff courtesy and sales efforts (Kurtovic & Hasimbegovic, 2015). These are only a few examples of mystery calling's advantages in affordability, flexibility, and effectiveness.

Wide Applicability

Companies have varying data needs, customer bases, market positions, and regulations. This is not necessarily a problem for mystery calling studies. Mystery calling has wide applicability across research contexts, ranging from car dealerships in Poland and the Czech Republic (Hys et al., 2017) to luxury hotels in Macau (Wan, 2010). These research contexts may include sensitive populations, products, and services, such as finance and healthcare. The wide applicability of mystery calling may also help UX researchers pursue a battery of research objectives: evaluating interactive voice response (IVR) systems, assessing how well a call center accommodates users with different abilities and speech patterns, understanding how well an organization's phone support integrates with other channels (such as in-person, app, and website), testing the consistency of information across a customer journey, carrying out error-

recovery testing, and gauging the effectiveness of a team's complaint handling, to name only a few. Mystery calling may thus promote UX research in work domains, countering criticism that the profession has become too focused on leisure domains (Caglar et al., 2022).

Cons of Mystery Calling

UX researchers must also consider the cons of mystery calling before deploying this method. The cons include medium bias, rigidity, standardization, and participant distrust, all of which can negatively impact data quality and subsequent decision-making.

Medium Bias

As the name suggests, mystery calling collects telephonic data. It often misses other data that matters during an interaction, such as nonverbal cues, documentation practices, and facility ambiance. Furthermore, staff may provide less information over the telephone than they would face-to-face, especially if the mystery call comes during a busy time or broaches a touchy subject. To improve data quality, UX researchers may need to conduct mystery visits, mystery emails, or mystery faxes in addition to mystery calls (Rady & Wahab, 2019).

Like all UX research, the findings from mystery calling depend on the sample, which may raise significant challenges if UX researchers rely on public directories. Telephone numbers may be inaccurate due to misinformation, company relocations, or participant retirements (Corbisiero et al., 2023). In addition, researchers may encounter long wait times and unanswered calls (Corbisiero et al., 2023), along with answering machines and unexpected disconnections (Kunow et al., 2021). These challenges may affect the sample and subsequent findings.

Rigidity

Mystery calls can become rigid in two ways. First, the research protocol may suppress the spontaneity that characterizes human conversation, making the data a poor reflection of business as usual. Second, mystery callers are anonymous, which limits the customer journeys they can trace. Specific to healthcare settings, Kunow et al. (2021) recommend a focus on products rather than on symptoms, and Kurtovic and Hasimbegovic (2015) evaluated basic banking tasks only: opening a checking account, requesting a new credit card, and inquiring about non-purpose loans. The study did not, and perhaps could not, evaluate more complex transactions.

Standardization

Standardization can become a challenge when a protocol includes subjective or open-ended questions. For instance, although their study was insightful, one team answered questions like these: "Did the employee kindly greet you?" "Did the employee listen to you carefully during your inquiry?" "Was the voice of the employee understandable and audible?" The terms "kindly," "carefully," "understandable," and "audible" are so open to interpretation that they may lead to inter- and intra-rater variability (Kurtovic & Hasimbegovic, 2015).

Further, the characteristics of the mystery callers may shape how participants respond. Investigating some of these effects, Wilkinson et al. (2017) compared how pharmacy workers reacted to three categories of mystery callers when asked about emergency contraception: female adult physicians, adolescent females, and adolescent males. The researchers found no significant differences when the mystery callers asked about the same-day availability of a contraception product ($p = 0.34$). However, compared to the adolescents, the physicians were significantly more likely to speak with a pharmacist, be transferred to a pharmacist, or be placed on hold ($p < 0.01$). Compared to adolescent females, adolescent males were more likely to be told, incorrectly, that they could not obtain emergency contraception because of their age ($p < 0.01$) (Wilkinson et al., 2018). Additional reactive effects may result from the characteristics of staff who answer the telephone (Lungfiel et al., 2023) or the location of the research site, urban or rural (Lilja et al., 2018; see also Giacomelli & Tonello, 2015).

Distrust

Deception can harm interpersonal and organizational relationships. In particular, deception may decrease liking, instill negative feelings, and trigger retaliation. It may also damage trust,

sometimes beyond repair (for a summary, see Levine & Schweitzer, 2015). Because mystery calling involves deception, albeit minor, this method may go over poorly with staff in debriefings. Participants may view the mystery calls as managerial surveillance or punishment. They may even perceive the mystery calls as threatening (Wady & Wahab, 2019). Although some research has suggested that "deception, per se, does surprisingly little to undermine trust behavior in the trust game," especially when the deception comes across as benevolent (Levine & Schweitzer, 2015, p. 102), but positive perception cannot be guaranteed.

Additional consequences may result from audio recording, which is not essential for mystery calling. Audio recording is legal in some jurisdictions, including most of the United States (Dickson et al., 2016), which requires one-party consent. This means that only one person on the call, who may be the researcher, needs to consent to the recording. Other jurisdictions require two-party consent, such as France and Germany. Complexities can arise over the definition of consent, which may be implied rather than explicit, over the reasonable expectation of privacy, and over differences between local and national laws. However, even where audio recording is permitted, participants may have a strong negative reaction when they are asked to consent to the recording, or when they are debriefed later.

This summary of pros and cons can help UX researchers determine whether to apply mystery calling in their own projects. Yet, the telephone, though familiar, is not a stable technology. As the telephone continues to evolve, mystery calling as a method will adapt. Much of the innovation in the coming years will likely result from AI.

Impact of AI

Sparse literature has tracked how mystery calling is changing amid the rapid proliferation of AI tools. In this section, we scan the horizon and suggest that, as in other areas of UX, AI will automate parts of data collection, data analysis, project management, and research reporting. To address their respective business needs, UX researchers will need to account for machine-machine, not only human-machine, interactions.

Data Collection

Data collection in mystery calling will transform as more UX researchers experiment with speech synthesis, which deploys AI-generated voices. To personalize a mystery call, researchers can select ready-made voices that differ by sex, accent, dialect, and other individual characteristics that may influence how participants respond. Researchers can further refine an AI voice through speech synthesis markup language, a derivative of XML, to refine pitch, speed, volume, and pauses. Some AI voices are text-to-speech (TTS), which may require typing in real time, depending on the research protocol. TTS has traditionally involved computational processes of articulatory synthesis, formant synthesis, concatenative speech synthesis, and statistical parametric techniques derived from a hidden Markov model; more recent developments involve deep learning. For details on these approaches, see Kaur and Singh (2023).

Other AI voices are based on voice conversion (VC), which is "the task of making a speech utterance from a *source* speaker sound like it came from a *target* speaker while keeping the linguistic content unchanged" (Triantafyllopoulous et al., 2023, p. 1361). As Triantafyllopoulous et al. (2023) explain, early approaches to VS employed articulatory synthesis. More recent developments have employed Gaussian mixture models and exemplar-based frameworks using nonnative matrix factorizations.

TTS and VC are only two systems for speech synthesis, and they are not mutually exclusive. They can run in combination, as Luong and Yamagishi (2020) illustrated. These researchers designed and validated a tool dubbed NAUTILUS, which switches between TTS and VC, a strategy that has become more common. AI tools like these can lighten the burden of data collection. If well-designed, AI tools may provide a high degree of automation by following protocols under human supervision.

AI tools can also expand data collection in mystery calling to use additional languages, even ones that the UX researchers do not speak well themselves. Samsung® (2023) introduced Live Translate, which does what its name promises: telephonically translates speech real time. A mystery caller could speak English into a microphone, and a participant in, say, Seoul could hear the mystery caller's voice in Korean. Among other advantages, these capabilities can

expand the vignettes that UX researchers deploy as mystery calls (see Corbisiero et al., 2023). As needed, UX researchers can update the vignettes to capture a wider sample of customer voices, needs, interests, and values.

Many companies already use machines to answer telephones. Yet, AI tools can do more than play recordings, prompt button presses, and connect the caller with a staff member if desired. They can interact with callers intelligently. A longstanding research area in telecommunications (Magedanz, 1995), intelligent agents are becoming more accessible and affordable. UX researchers can expect to encounter an increasing number of intelligent agents during mystery calls, and the current industry standard may be Lucy and Sam by Curious Thing™ (2024). In time, mystery calling may involve less human-machine interaction but more machine-machine interaction. Further innovations that can enhance data collection, and perhaps data analysis, in mystery calling will appear at industry expos, such as the Mobile World Congress.

Importantly, AI innovations in data collection raise significant questions in mystery caller studies. At some point, UX researchers may find it difficult to distinguish between human and machine responses, especially as speech synthesis advances. AI tools may provide more consistent responses than humans, making it a challenge to evaluate variations in service. Moreover, AI tools may also become sophisticated enough to filter out, or adapt to, mystery calls, potentially biasing a sample. These are only a few areas of concern.

Data Analysis

Methods such as natural language processing (NLP), a component of AI, can power the analysis of mystery caller data. NLP uses statistics and machine learning to surface patterns in large data sets and then identify important relationships. Some of the most common uses of AI are marking parts of speech, extracting objects upon request, and disambiguating the data. Perhaps especially valuable for mystery caller studies, NLP can cluster words and phrases from transcripts into meaningful topics (see Dredze et al., 2010), as well as summarize the content of audio files (Prowal et al., 2022). NLP is not a new development; however, according to Abdusalomovna (2023), the technology has made major strides in the last few years. Future innovations in NLP will pivot from descriptive to predictive analysis, enabling UX researchers to better anticipate participant questions and informational needs (Wang, 2022).

AI tools, including NLP, are not without shortcomings. As Silberztein (2024) explained, NLP applications rely on training corpora, which teach the underlying algorithms how to analyze data. However, these corpora often contain many mistakes and may bear little resemblance to the texts later under analysis. Beyond basic errors in data processing, questions arise over how effectively AI tools can parse data that contains industry-specific information, non-standard language, and nuanced emotion. The results may be unempirical and unreliable.

Project Management

In mystery caller studies, UX researchers can facilitate project management through AI tools such as virtual assistants, which can learn from the researchers' behavior and message and schedule on their behalf. These functions may involve analyzing multiple calendars, proposing new meeting times that avoid conflicts, and arranging project tasks in ways that maximize efficiency (Baek et al., 2023). An early but marked example is SRI International's Cognitive Agent that Learns and Organizes (CALO). Explained Brachman (2006):

> *What CALO is trying to emulate is not a Ph.D.-level physicist, but rather, the seemingly mundane tracking, learning, and reminding aspects of a good secretary who can adapt to real-world circumstances, improve over time, become personalized to the person he or she is supporting, and take into account the many small things that make everyday life challenging. (p. 29)*

More of these virtual assistants will become available in the coming years, supporting the project management of mystery caller studies.

However, virtual assistants may also increase project costs, integrate poorly with existing systems, make errors, or even become a crutch that UX researchers overly depend on.

Research Reporting

Generative AI may assist with research reports, allowing UX researchers to document their findings from mystery caller studies faster. Recently, ChatGPT received credit as the lead author

of a publication in *Oncoscience* (Transformer & Zhavoronkov, 2022). Of course, crediting an AI tool with authorship is controversial, especially because it cannot detect its own fabrications, falsifications, or "hallucinations" (Emsley, 2023). Subsequently, UX researchers will need to tread with care, vetting output from AI tools.

This is only a cross-section of the AI tools currently on the market or under development. Trends suggest that they will continue to increase in power and prevalence, impacting mystery caller methods.

Recommendations

The documented pros and cons of mystery calling, together with the impacts of AI, should inform best practices. Building on prior sections, we recommend a few best practices for UX researchers who deploy mystery calling in their own work.

Pre-Test

Before collecting data, usability test the mystery caller protocol and modify it as needed (van Hoof et al., 2014). This pre-testing may involve a small sample of the target population, generating data that should be excluded if the protocol changes. Another option is to sample outside of the target population, for instance, for those who live in a different city, state, or region.

Standardize Data Collection Forms

Create a standardized form for mystery callers to fill out, focusing on relevant metrics or evaluation criteria. Depending on the research objectives, the form may provide yes/no questions, Likert scales, or blank spaces for additional call notes. Save each form in a secure location in case the team needs to review it later. For a detailed example, see the supplement to Egan et al. (2019).

Forgo Audio Recording

We advise against audio recording, even where it is permitted under one-party consent. Limiting data collection to written notes avoids legal and ethical complexities, reduces the capture of extraneous information, and can reduce consequences with participants. Moreover, a well-designed data collection form makes audio recordings unnecessary.

Orientation

For additional standardization, hold a pre-launch orientation for team members on the study's purposes and procedures. The orientation may provide practice in following the protocol, taking notes or other recordings, and storing data, among other topics. After officially launching the study, hold regular meetings to discuss challenges, such as technical glitches or unanticipated questions from participants.

Call Back

Participants may not answer the telephone at first, especially during busy hours. For a more comprehensive sample, call two (Corbisiero et al., 2023) or three times (Egan et al., 2019) before marking the participants as non-responsive.

Stay Flexible

Because human conversation is spontaneous, not every mystery call will go according to plan. To avoid rigidity, consider using a semi-structured protocol that allows UX researchers to deviate as needed, and then civilly and subtly redirect the conversation to the study objectives. Vignettes, as Corbisiero et al. (2023) described, can provide character backstories that can help UX researchers address unexpected questions. Regardless of the design, match the energy of the participant. Doing so may help put them at ease.

Work Expediently

Because data from mystery calls is time-sensitive, expediency is key. The appropriate speed, however, depends on the study objectives. A mystery caller study may reasonably take place over a few days (Corbisiero et al., 2023; Kurtovic & Hasimbegovic, 2015), a few weeks (Dickson et al., 2018; Lungfiel et al., 2023), a few months (Ditmars et al., 2019), or a few semesters (Kunow et al., 2021).

Apply the Four-Eyes Principle

Data uploads can cause errors. For instance, UX researchers may mistranslate their call notes into a shared table that the team is analyzing together. The four-eyes principle provides a form of quality assurance, as Kunow et al. (2021) suggested. To apply it, have a second researcher—who is competent and independent—double-check the work of team members. This may involve auditing a sample of data or undertaking a complete review.

Create a Codebook

A codebook is a document that describes each variable, enabling UX researchers to better understand the data. The codebook may include variable names and labels, values or codes assigned to them, missing data codes, special instructions on data usage, and other relevant information (Bélisle & Joseph, 2015). For instance, the codebook may additionally include notes on the data collection methods, units of measurement, formatting specifications, and the version number and date of last update. As a result, a codebook helps UX researchers analyze the data in a standardized way and, when desired, test for inter-rater reliability. For an extensive example, see the supplement to Egan et al. (2019), available at https://dataverse.unc.edu/dataverse/naloxoneinpharmacy.

Open the Black Box

AI tools have considerable potential to enhance data collection, data analysis, project management, and reporting of mystery caller studies. Proceed with some caution, however, because AI tools are often a black box. A term that originated in electronics, a black box is an opaque system that produces output through processes only somewhat known to the users (see Castelvecchi, 2016). To compensate, collaborate with specialists who can parse the inner workings of AI tools, assess strengths and weaknesses, recommend which AI tools (if any) to deploy in a mystery caller study, and monitor the data quality. At a minimum, this means that AI tools, like a protocol, should undergo pre-tests before implementation. UX researchers may find, for instance, that AI-generated voices are still too robotic to interact meaningfully with participants, or that AI-generated voices do not sound like local customers. Since the costs for mystery caller studies are minimal, we recommend that UX researchers make the calls themselves.

Be Transparent

Be transparent in at least three ways. First, debrief the participants, revealing the minor deception and study purposes, which may be required by the review board overseeing a mystery caller study. Debriefing is also considered a best practice in research ethics (Market Research Society, 2024).

Kunow et al. (2021) planned to debrief the participants before the mystery calls. Depending on the study objectives, other researchers may prefer to debrief after the mystery calls are complete. That way, they can share results, invite comments and questions, and protect the integrity of data collection. If participants learn of the mystery caller study before or during data collection, social desirability bias and the Hawthorne effect may arise, undermining the (near) authenticity of the data.

Second, acknowledge the use of AI tools, including in the production of the final research report. The Committee on Publication Ethics (COPE) (2003) does not recognize AI tools as authors, and for good reason.

Third, while companies may issue their own standards for research reporting, consider adapting the STROBE guidelines (Equator Network, 2023). These guidelines provide a checklist of useful information that can contextualize a mystery caller study, such as details on the methods, results, and implications.

Weigh the Ethics Carefully

Whether assisted by AI tools or not, mystery calling requires a careful balance of ethical principles. To start, seek to minimize risk, such as by omitting from call notes the precise date of the mystery call, the time of the mystery call, and the names of the participants. Meanwhile, find ways to maximize benefits, such as by sharing results with participants and acknowledging limitations of the methods. For instance, mystery calls may capture only a portion of a service interaction, missing non-verbal cues, documentation, or perhaps more extensive conversation

that employees would hold in-person. For a more complete perspective, supplement with additional research methods as needed.

UX researchers should review ethics guidelines from ESOMAR and the Global Research Business Network (n.d.), the Market Research Society (2024), and the Mystery Shopping Providers Association (2011) as they plan mystery caller studies.

Conclusion

In this article, we have offered a critical discussion of mystery calling for UX research. We believe that mystery calling may indeed provide a low-cost, high-yield method for UX researchers across industries given the pros and cons and potential impacts of AI.

We invite readers of this journal to explore further, opening new avenues of inquiry and professional practice.

Tips for Usability Practitioners

- Craft specific and realistic vignettes that mystery callers can use to test various aspects of your product or service. These scenarios should reflect typical and atypical user interactions to ensure well-rounded usability testing. As your company and customer interactions evolve, regularly update the vignettes to keep them relevant.
- Be aware of medium bias. Recognize that mystery calling often misses non-verbal cues and may provide less information than face-to-face interactions. Consider complementing mystery calls with other methods as needed.
- For post-calls, aggregate and analyze the data to identify friction points. For large datasets, experiment with NLP and similar AI techniques. Vet any AI-generated results.

References

Abdusalomovna, T. D. (2023). Text mining. *European Journal of Interdisciplinary Research and Development*, *13*, 284–289.

Baek, S., Kim, J., Lee, J., & Lee, M. (2023). Implementation of a virtual assistant system based on deep multi-modal data integration. *Journal of Signal Processing Systems*, 1–11. https://doi.org/10.1007/s11265-022-01829-5

Bergen, N., & Labonté, R. (2020). "Everything is perfect, and we have no problems": Detecting and limiting social desirability bias in qualitative research. *Qualitative Health Research*, *30*(5), 783–792.

Brachman, R. J. (2006). AI more than the sum of its parts. *AI Magazine*, *27*(4), 19–19.

Bucher, A., Schenk, B., & Schwabe. (2023). When learning turns to surveillance—Using pedagogical agents in organizations. *Proceedings of the 56th Hawaii International Conference on System Sciences*, *2023*, 247–256.

Caglar, P.S., Roto, V., & Vainio, T. (2022). User experience research in the work context: Maps, gaps and agenda. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW1), 1–28.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News*, *538*(7623), 20.

Committee on Publication Ethics (2023, February 13). *Authorship and AI tools: COPE position statement*. COPE. https://publicationethics.org/cope-position-statements/ai-author

Corbisiero, M. F., Tolbert, B., Sanches, M., Shelden, N., Hachicha, Y., Dao, H., & Muffly, T. M. (2023). Medicaid coverage and access to obstetrics and gynecology subspecialists: Findings from a national mystery caller study in the United States. *American Journal of Obstetrics and Gynecology*, *228*(6), 722–e1.

Curious Thing. (2024). *Never drop the ball/call with Lucy*. https://curiousthing.io/products/lucy-voice-ai-assistant-for-business

Da Silva, T.S., Silveira, M.S., Maurer, F., & Hellmann, T. (2012). User experience design and agile development: From theory to practice. *Journal of Software Engineering and Applications, 5*, 743–751.

Dickson, B., Mansfield, C., Guiahi, M., Allshouse, A.A., Borgelt, L.M., Sheeder, J., Silver, R.M., & Metz, T.D. Recommendations from cannabis dispensaries about first-trimester cannabis use. *Obstetrics & Gynecology*, *131*(6), 1031–1038. https://doi.org/10.1097/AOG.0000000000002619

Ditmars, L., Rafie, S., Kashou, G., Cleland, K., Bayer, L., & Wilkinson, T.A. (2019). Emergency contraception counseling in California community pharmacies: A mystery caller study. *Pharmacy*, *7*(2), 38.

Dredze, M., Jansen, A., Coppersmith, G., & Church, K. (2010, October). NLP on spoken documents without ASR. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 460–470).

Egan, K.L., Foster, S.E., Knudsen, A.N., Lee, J.G.L. (2019). Naloxone availability in retail pharmacies and neighborhood inequities in access. *American Journal of Preventive Medicine*, *58*(5), 699–702.

Emsley, R. (2023). ChatGPT: These are not hallucinations—They're fabrications and falsifications. *Schizophrenia*, *9*(1), 52.

ESOMAR / Global Business Research Network (GRBN). (n.d.). *ESOMAR / GRBN guidelines for researchers and clients involved in primary data collection*. ESOMAR.org

Equator Network. (2023, March 6). *The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies*. https://www.equator-network.org/reporting-guidelines/strobe/

Gardner-Bonneau, D. (2010). Is technology becoming more usable—or less—and with what consequences? *Journal of Usability Studies*, *5*(2), 46–49.

Giacomelli, S., & Tonello, M. (2015). Measuring the performance of local governments: Evidence from mystery calls. *Bank of Italy Occasional Paper*, (292).

Gravlee, E., Ramachandran, S., Cafer, A., Holmes, E., McGregor, J., Jordan, T., & Rosenthal, M. (2023). Naloxone accessibility under the state standing order across Mississippi. *JAMA Network Open*, *6*(7), e2321939–e2321939.

Hinderks, A., Mayo, F. J. D., Thomaschewski, J., & Escalona, M. J. (2022). Approaches to manage the user experience process in Agile software development: A systematic literature review. *Information and Software Technology*, *150*, 106957.

Hys, K., Hawrysz, L., Kozel, R., & Vilamová, Šárka. (2017). Application of mystery calling method in car dealerships—Polish-Czech research. In O. Dvouletý, M. Lukeš, & J. Mísař (Eds.), *Proceedings of the 5th international conference on innovation management, entrepreneurship, and sustainability (IMES 2017)* (pp. 324–337).

IBM. (2024, January 5). *4 eyed principle*. https://www.ibm.com/docs/en/b2b-integrator/6.1.1?topic=principle-4-eyed

Kaur, N., & Singh, P. (2023). Conventional and contemporary approaches used in text to speech synthesis: A review. *Artificial Intelligence Review*, *56*(7), 5837–5880.

Khan, T., Williams, M. A., & Pitts, M. J. (2016). Cross-cultural differences in automotive HMI design: A comparative study between UK and Indian users' design preferences. *Journal of Usability Studies*, *11*(2), 45–65.

Kunow, C., Bello, M.A., Diedrich, L., Eutin, L., Sonnenberg, Y., Wachtel, N., & Langer, B. (2021). A nationwide mystery caller evaluation of oral emergency contraception practices from German community pharmacies: An observational study protocol. *Healthcare, 9*(8), 945.

Kurtovic, E., & Hasimbegovic, A. (2015). Measuring customer service level in banking sector applying mystery calls method and its relation to the HR department. *The Business & Management Review*, *6*(3), 1–11.

Lamm, L., & Wolff, C. (2021). GCS: A quick and dirty guideline compliance scale. *Journal of Usability Studies*, *16*(3), 179–202.

Levine, E. E., & Schweitzer, M. E. (2015). Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes*, *126*, 88–106.

Lewis, J. R., & Sauro, J. (2021). Comparison of select-all-that-apply items with yes/no forced choice items. *Journal of Usability Studies*, *17*(1), 21–30.

Lungfiel, G., Mandlmeier, F., Kunow, C., & Langer, B. (2023). Oral emergency contraception practices of community pharmacies: A mystery caller study in the capital of Germany, Berlin. *Journal of Pharmaceutical Policy and Practice, 16*(1), 68.

Luong, & Yamagishi, J. (2020). Nautilus: A versatile voice cloning system. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 2967–2981.

Macefield, R. (2007). Usability studies and the Hawthorne Effect. *Journal of Usability Studies*, *2*(3), 145–154.

Market Research Society (MRS). (2004). *MRS guidelines for conducting mystery shopping*. https://www.mrs.org.uk/standards/mrs-guideline-for-conducting-mystery-shopping

Magedanz, T. (1995). On the impacts of intelligent agent concepts on future telecommunication environments. In *Bringing Telecommunication Services to the People—IS&N'95: Third International Conference on Intelligence in Broadband Services and Networks Heraklion, Crete, Greece, October 16–19, 1995 Proceedings 3* (pp. 394–414). Springer Berlin Heidelberg.

Mystery Shopping Providers Association. (2011). *Guidelines for mystery shopping*. mspa-ea.org

Pollack, C.E., Ross, M.E., Armstrong, K., Branas, C.C., Rhodes, K.V., Bekelman, J.E., Wentz, A., Stillson, C., Radharkrishnan, A., Oyeniran, E., & Grande, D. (2016). Using a mystery-caller approach to examine access to prostate cancer care in Philadelphia. *PLOS One*, *11*(10), e0164411.

Porwal, K., Srivastava, H., Gupta, R., Pratap Mall, S., & Gupta, N. (2022). Video transcription and summarization using NLP. *Proceedings of the Advancement in Electronics & Communication Engineering*.

Qualcomm. (n.d.). *Snapdragon 8 Gen 3 Mobile Platform*. https://www.qualcomm.com/products/mobile/snapdragon/smartphones/snapdragon-8-series-mobile-platforms/snapdragon-8-gen-3-mobile-platform

Rady, A., & Wahab, H. A. (2019). Mystery shopper as a tool to measure staff performance in travel agencies. *Minia Journal of Tourism and Hospitality Research MJTHR*, *8*(1), 1–28.

Robinson, J., & Lanius, C. (2018). A geographic and disciplinary of UX empirical research since 2000. *SIGDOC '18*. Milwaukee, WI: USA.

Rummel, B. (2014). Probability plotting: A tool for analyzing task completion times. *Journal of Usability Studies*, *9*(4), 152–172.

Samsung. (2024, January 30). *How to use Live translate for phone calls on the Galaxy S24*. https://www.samsung.com/latin_en/support/mobile-devices/how-to-use-live-translate-for-phone-calls-on-the-galaxy-s24/#:~:text=a%20phone%20call.-,Step%201.,Tap%20on%20Live%20translate

Scerbo, M. W. (2023). Can artificial intelligence be my coauthor? *Simulation in Healthcare*, *18*(4), 215–218.

Silberztein, M. (2024). The limitations of corpus-based methods in NLP. In M. Silberztein (Ed.), *Linguistic resources for natural language processing: On the necessity of using linguistic methods to develop NLP software* (pp. 3–24). Cham: Springer Nature Switzerland.

Transformer, C. G. P. T., & Zhavoronkov, A. (2022). Rapamycin in the context of Pascal's Wager: Generative pre-trained transformer perspective. *Oncoscience*, *9*, 82.

Triantafyllopoulos, A., Schuller, B. W., İymen, G., Sezgin, M., He, X., Yang, Z., & Tao, J. (2023). An overview of affective speech synthesis and conversion in the deep learning era. *Proceedings of IEEE, 111*(10), 1355–1381.

Tullis, T.S. (2009). Tips for usability professionals in a down economy. *Journal of Usability Studies, 4*(2), 60–69.

Turner, H. (2012). Mystery shopping. In M. van Hamersveld & C. de Bont (Eds.), *Market research handbook* (pp. 333–346). ESOMAR World Research Publication.

Van Hoof, J. J., Van Den Wildenberg, E., & De Bruijn, D. (2014). Compliance with legal age restrictions on adolescent alcohol sales for alcohol home delivery services (AHDS). *Journal of Child & Adolescent Substance Abuse*, *23*(6), 359–361.

Verstraete, J., & Verhaeghe, P. P. (2020). Ethnic discrimination upon request? Real estate agents' strategies for discriminatory questions of clients. *Journal of Housing and the Built Environment*, *35*(3), 703–721.

Wan, Y.K.P. (2010). Promoting hotel service quality through managing reservationist call-handling performance. *Journal of Quality Assurance in Hospitality and Tourism*, *11*(3), 199–218.

Wang, Y. (2022). Using machine learning and natural language processing to analyze library chat reference transcripts. *Information Technology and Libraries*, *41*(3).

Wilkinson, T. A., Rafie, S., Clark, P. D., Carroll, A. E., & Miller, E. (2018). Evaluating community pharmacy responses about levonorgestrel emergency contraception by mystery caller characteristics. *Journal of Adolescent Health*, *63*(1), 32–36.

About the Authors

Michael J. Madson
Michael is an assistant professor in the user experience and technical communication programs at Arizona State University. His current research focuses on drug safety, especially opioids and cannabis.

Yoshita Gade
Yoshita came to UX from architecture. She has worked as a product designer at Rythmos and Global Launch ASU. She recently earned her master's degree in UX from Arizona State University.

Unnati Srivastava
Unnati has a background in computer science. She has a wide range of experience working with B2C, B2B, and SaaS clients. She also earned her master's degree in UX from Arizona State University.