An Evaluation of Temporal and Categorical Uncertainty on Timelines: A Case Study in Human Activity Recall Visualizations

Veronika Potter[*] Ha Le Uzma Haque Syeda Stephen Intille[†] Michelle A. Borkin

Khoury College of Computer Sciences, Northeastern University

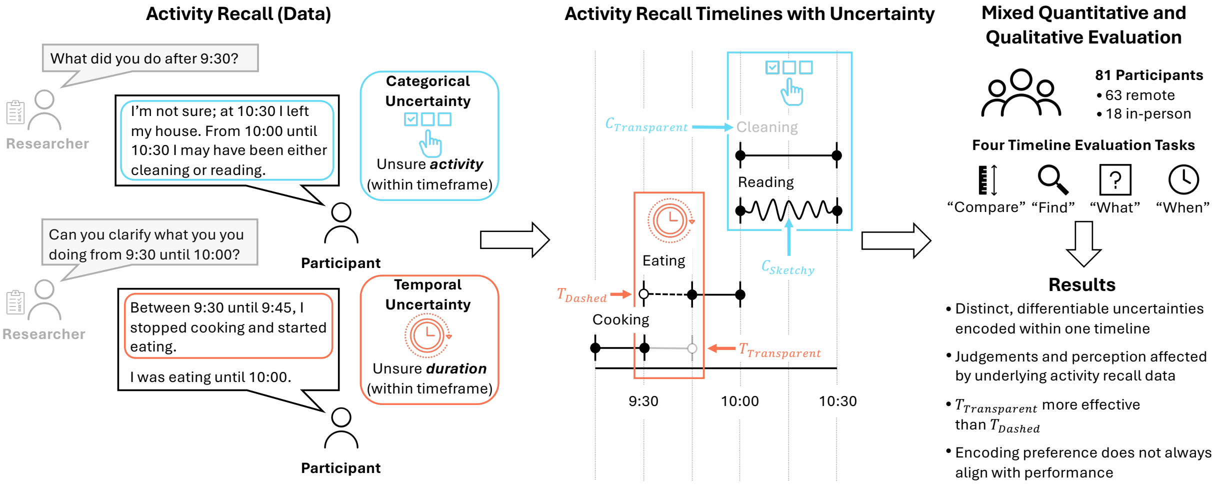
Activity Recall (Data)
What did you do after 9:30?
Researcher
I'm not sure; at 10:30 I left my house. From 10:00 until 10:30 I may have been either cleaning or reading.
Participant
Categorical Uncertainty
Unsure *activity* (within timeframe)
Can you clarify what you you doing from 9:30 until 10:00?
Researcher
Between 9:30 until 9:45, I stopped cooking and started eating.
I was eating until 10:00.
Participant
Temporal Uncertainty
Unsure *duration* (within timeframe)
Activity Recall Timelines with Uncertainty
$C_{Transparent}$
Cleaning
Reading
$C_{Sketchy}$
Eating
T_{Dashed}
Cooking
$T_{Transparent}$
9:30
10:00
10:30
Mixed Quantitative and Qualitative Evaluation
81 Participants
• 63 remote
• 18 in-person
Four Timeline Evaluation Tasks
"Compare" "Find" "What" "When"
Results
• Distinct, differentiable uncertainties encoded within one timeline
• Judgements and perception affected by underlying activity recall data
• $T_{Transparent}$ more effective than T_{Dashed}
• Encoding preference does not always align with performance

Figure 1: Activity recall timelines (ARTs) can encode uncertainty in activity recalls for better sensemaking while conveying the data's underlying validity. We conducted a mixed-method user study to evaluate different temporal and categorical uncertainty encodings in ARTs.

Abstract

Encoding uncertainty in timelines can provide more precise and informative visualizations (e.g., visual representations of unsure times or locations in event planning timelines). To evaluate the effectiveness of different temporal and categorical uncertainty representations on timelines, we conducted a mixed-methods user study with 81 participants on uncertainty in activity recall timelines (ARTs). We find that participants' accuracy is better when temporal uncertainty is encoded using transparency instead of dashing, and that a participant's visual encoding preference does not always align with their performance (e.g., they performed better with a less-preferred visual encoding technique). Additionally, qualitative findings show that existing biases of an individual alter their interpretation of ARTs. A copy of our study materials is available at `https://osf.io/98p6m/`.

Index Terms: Timelines, Uncertainty Visualization, Evaluation Study.

1 Introduction

A labeled list of an individual's daily activities can be used to monitor physical activity habits, sleep patterns, and sedentary behavior levels. Additionally, accurate daily activity depictions can be used for changing behavioral patterns [25], self-reflection [4], understanding a population's time usage (e.g., the American Time Use Survey [1]), and training human activity recognition models. However, collecting an accurate depiction of an individual's activities is difficult and is oftentimes done through a "self-report activity recall:" an individual reporting their activities via handwritten or digitized logs (e.g., [18]), conversations with researchers, frequent surveys (e.g., [15, 26]), or custom mobile or web applications (e.g., [23]). Because self-report activity recalls are issued after the fact and may ask about a specific point in time (e.g., "what were you doing between [time t_1] and [time t_2]?"), they are subject to recall bias [26, 23] that may affect the duration (temporal uncertainty) and/or the type (categorical uncertainty) of a recalled activity. Although activity recalls are often interpreted from a textural or tabular format, activity recall timelines (ARTs) allow for better pattern identification and sensemaking of the underlying data [20, 8, 29, 13, 27]. To convey the validity of each activity in the recall, ARTs must encode the temporal or categorical uncertainty present in the original activity recall.

ARTs (shown in Fig. 2) are only one example in which encoding uncertainty may yield more precise and informative visualizations. Travel planning timelines could encode potential delays or changing destinations, and historic timelines could encode theorized but unproven events or people. To study uncertainty encodings in timelines, we investigate four research questions in the context of ARTs:

RQ1: *Is a transparent or dashed time frame more effective in conveying **temporal** uncertainty?*
RQ2: *Is a transparent activity label or sketchy time frame more effective in conveying **categorical** uncertainty?*
RQ3: *How does ART comprehension vary by task and timeline density?*
RQ4: *Which existing biases or mental models impact ART perception?*

We conducted a mixed-method evaluation study for visual uncertainty encodings on ARTs with 81 participants. Each participant completed an online survey (quantitative data). In-person participants also participated in a semi-structured interview (qualitative data). We make two contributions:

1) The results, and accompanying discussion, of our evaluation study in which we compare the effectiveness and human preference for different uncertainty encoding techniques on timelines in the context of ARTs.
2) Proposed design recommendations for timelines that encode temporal and categorical uncertainty.

2 Related Literature

Uncertainty representations visually represent uncertainty in the underlying data and have heavily focused on data that reflects the probability of something occurring (e.g., transit arrival time [12] or a hurricane's

[*] Corresponding author: potter.v@northeastern.edu.
[†] Dual appointment at Bouvé College of Health Sciences.

trajectory [22]). Visualizations have encoded this probability by changing the granularity of the underlying data [6]; using probability distributions, both discrete [9, 12] and continuous [21]; and including error bars, fuzziness, and transparency [14, 21]. Uncertain linear data has been encoded using dashing, grayscale, and sketchy lines [7]. **Uncertainty encodings in timeline visualizations** are an open area of research [10, 24]; ours is the first work to evaluate sketchiness in timelines. Planning tools incorporating temporal uncertainty have been introduced and evaluated, but the encodings themselves (interval bars and transparency [3] as well as color, shape, and comparative length [16]) had not been independently evaluated. In a user study evaluating multiple *temporal* uncertainty encodings, ambiguation was found to be an effective communicator of temporal uncertainty but the *ideal* representation was task-dependent, especially when underlying probability needed to be communicated [14]. To the best of our knowledge, no literature proposes or evaluates visual encodings for categorical uncertainty in timeline visualizations. Further, no work evaluates both temporal and categorical uncertainty within the same timeline. **Evaluating timeline visualizations** often takes into account both participant performance (quantitative) and perception (qualitative). To evaluate timelines shape, Di Bartolomeo et al. proposed a framework of four timeline evaluation tasks: "Compare," "Find," "What," and "When;" they found linear timelines were best and that task complexity may alter performance [11]. Temporal uncertainty in timelines has been evaluated using tasks like "probabilities" (probabilistic estimation) and "start/stop" ("When" tasks) [14]. These quantitative evaluations used metrics such as participant accuracy, response time, and confidence [11, 14]. To extend on prior research, our work evaluates different binary temporal and categorical uncertainty visualization both quantitatively, using similar metrics as [14] and [11], and qualitatively to provide insight on the impacts of underlying data/application on understanding and use of uncertainty in timelines.

3 Methodology

3.1 Stimuli Design

We designed the ARTs using two rounds of an iterative design process; each round consisted of internal brainstorming, refining designs with visualization experts, and pilot testing with non-data visualization experts. During each brainstorming phase, we proposed or modified uncertainty representations affecting the endpoint glyphs, time frames, and activity labels. During pilot testing, we qualitatively evaluated every representation as encoding each categorical and temporal uncertainty. Ultimately, based primarily on pilot testing feedback, we chose four uncertainty encodings to evaluate. Example stimuli are shown in Fig. 2.

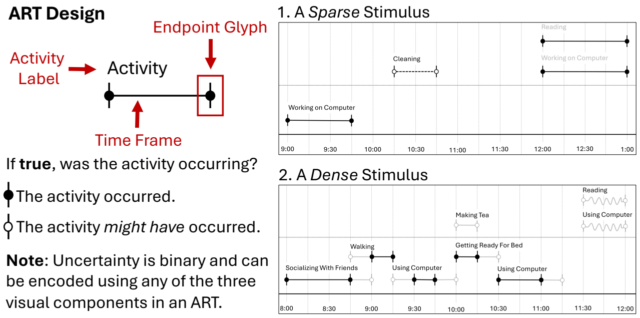

Figure 2: Design of ARTs and example stimuli. Left: visual components in ARTs. Right: a sparse (top) and dense (bottom) stimulus.

We evaluated two encodings for temporal uncertainty: $T_{Transparent}$ and T_{Dashed} (Fig. 1). We chose $T_{Transparent}$ because ambiguation has performed well to express temporal uncertainty [14] and our pilot testers qualitatively preferred $T_{Transparent}$ and T_{Dashed} to other options. Both were visually encoded by changing the time frame shape (i.e., stroke style) and endpoint glyphs' closure (i.e., open/filled), and followed the intuition that temporal uncertainty impacts the activity duration, thus should visually be represented by the time frame and endpoints. Additionally, we evaluated two encodings for categorical uncertainty: $C_{Transparent}$ and $C_{Sketchy}$ (Fig. 1). We chose $C_{Transparent}$ as it was our pilot testers most preferred *intuitive* encoding (categorical uncertainty impacts the activity type, visually represented by the activity label). However, because text is interpreted differently than other visual elements [28] and some pilot testers missed the $C_{Transparent}$ encoding at a glance, we chose to include $C_{Sketchy}$ (a highly preferred encoding on the time frame that used a unique channel) as an encoding for categorical uncertainty but did not alter the endpoints because the uncertainty is not associated to time. To encode categorical uncertainty, $C_{Transparent}$ changes the transparency of the activity label and $C_{Sketchy}$ changes the shape of the time frame. Because a recalled activity may be both categorically and temporally uncertain we chose encodings that used unique visual channels so they could be visually layered.

We created twenty stimuli by modifying an existing activity dataset [2]: sixteen containing both temporal and categorical uncertainty and four without (control stimuli). Half of the stimuli (10) are *sparse* (contain less than or equal to five activities) and the remainder (10) are *dense* (contain more than five activities).

3.2 Study Protocol

We conducted a mixed methods evaluation study with both ***in-person*** and ***remote*** participants balancing the need for both statistical power and high-quality qualitative data. Participants answered questions during the **online survey** about both temporal uncertainty encodings (within) and only one categorical uncertainty encoding (between).We assigned study conditions, evaluating $C_{Transparent}$ or $C_{Sketchy}$, in a round-robin fashion regardless of participation modality.

Participants received 68 questions in the **online survey**: 4 identical control questions + 2 conditions * 2 densities * (16 task-based questions). Within each study condition, the stimuli order was counterbalanced using a Latin square design to reduce ordering effects. We derived each question from the four evaluation tasks used to evaluate timeline shape [11] ("Compare," "Find," "What," and "When"). We modified each task to be realistic for ARTs (Table 1). Although we designed the **online study** to take our pilot testers no longer than 45 min in one sitting, we did not limit completion time and participants were not penalized if they took a break. Additionally, we divided the control questions equally between the two conditions to ensure participant comprehension over time.

Table 1: ART evaluation tasks (following the framework of Di Bartolomeo et al. [11]), descriptions, and example questions.

Task	Description	Example Question
↕ "Compare"	Locate the shortest/longest recalled activity or explain the relationship between two activities in the timeline.	Which activity (or activities) could this individual have spent the least time doing?
🔍 "Find"	Find the total duration of an activity (that may have multiple bouts) in the timeline.	How many minutes could this individual have been *Using Computer*?
? "What"	Identify all activities at a specific time or within a specific time frame.	Which activity (or activities) could this individual have been doing at 8:00?
🕒 "When"	Find the start, stop, or time frame of a specific activity.	When could this individual have been *Walking*?

Remote participants completed the online survey via Prolific (https://www.prolific.com/). ***In-person*** participants completed the online survey alongside a researcher in a *think aloud* session to explain their thought process, responses, and impressions during the **online survey**. After completing the online survey, ***in-person*** participants were sequentially shown additional stimuli containing all four uncertainty encodings during a semi-structured interview. With ***in-person*** participant consent, audio was recorded and transcribed. ***Remote*** participants received $10 USD and ***in-person*** participants received $15 USD in gift cards as compensation.

Northeastern University's IRB approved this research and all participants gave consent before participating in our study. We preregistered our study procedures and analysis plan (https://osf.io/98p6m/).

3.3 Participants and Recruitment

We conducted a power analysis, using pilot testing data, prior to the study, which yielded a sample size of 68 participants. We aimed to recruit 76 participants (assuming 10% unusable data). Participants were at least 18 years old, not visually impaired, and resided in the United States at the time of participation. We recruited participants via Prolific (***remote***) as well as via flyers and email (***in-person***).

We recruited 81 participants, 18 (22%) ***in-person*** and 63 (78%) ***remote***. By self-reported demographics, our participants were approximately split by sex (39 (48%) were female at birth), aged 18-55+ (40 (49%) were ages 25-34), racially diverse (34 (42%) were White/Caucasian), and highly educated (56 (69%) had at least an associate's degree).

3.4 Hypotheses

We tested four hypotheses to address **RQ1-3** (Sec. 1). The first two hypotheses investigate effective uncertainty encodings (**RQ1** and **RQ2**):

H1: *Participants are more accurate when transparency on the time frame, rather than dashing, encodes temporal uncertainty.*

H2: *Participants are more accurate when transparent activity labels, rather than sketchy time frames, encode categorical uncertainty.*

We hypothesized that participants are more accurate using $T_{Transparent}$ than T_{Dashed}, **H1**, because ambiguation, which we achieve through transparency, has effectively encoded temporal uncertainty [14]. Additionally, we hypothesized that participants perform better on $C_{Transparent}$ than $C_{Sketchy}$, **H2**, because, although there is no prior work evaluating categorical uncertainty encodings, our pilot testers qualitatively preferred $C_{Transparent}$ to $C_{Sketchy}$ saying it is *more intuitive* because it visually encodes categorical uncertainty on the activity label, the uncertain element in the recall.

Our last two hypotheses address **RQ3** and evaluate the effects of density and task on perception and understanding of ARTs:

H3: "Compare," "Find," and "When," tasks will be completed quicker on sparse *stimuli, regardless of uncertainty encodings.*

H4: "What" tasks will be more accurate than "Compare," "Find," and "When" tasks regardless of density and uncertainty encodings.

We hypothesized that tasks that require navigating the entire timeline ("Compare," "Find," and "When") will be quicker on sparse ARTs, **H3**, because fewer activites are visualized and more cluttered visualizations result in slower response times [19]. Also, we hypothesized that "What" tasks will be the most accurate, **H4**, because they are simpler, do not require navigating all the activities in the timeline, and evaluation task complexity may alter response time and accuracy [11].

3.5 Data Analysis and Exclusion of Data

Quantitative Analysis: We used data from the **online survey** (participant responses and response time) for hypothesis testing. Responses were graded as correct or not with no partial credit. We computed accuracy as the number of correct responses divided by the total number of questions in that condition.

We tested for normality and evaluated **H1**, **H2**, and **H3** using one-tailed t-tests because each hypothesis 1) evaluated a difference between two groups; 2) was apriori and unidirectional; and 3) was guided by prior work (in the case of **H1** and **H3**) or pilot testing feedback (in the case of **H2**) [17]. We did additional post-hoc analysis of confidence and response times using one-tailed t-tests. We evaluated **H4** with a repeated measures ANOVA test because it considered multiple design factors. We used the Benjamini and Hochberg procedure [5] to minimize false discoveries.

We excluded the following data from our analyses 1) a participant's responses if they answered any control question incorrectly (11 participants (14%)), 2) response times of ***in-person*** participants because they were asked to think aloud, 3) any questions where greater than 80% of participants answered "Not enough information" or skipped the question (eight "Compare" questions).

Qualitative Analysis: One author carefully read and coded the transcripts of ***in-person*** interviews inductively. Iteratively, as a group, we discussed and improved the codes. We derived themes after removing codes outside the scope of our RQs and merging similar codes.

4 Results

We statistically analyzed the online-survey responses of the 70 (87%) included participants to address **RQ1**-**RQ3** and thematically analyzed the transcripts of all 18 (22%) in-person participants to address **RQ4**. Unabridged results are reported in our supplemental material.

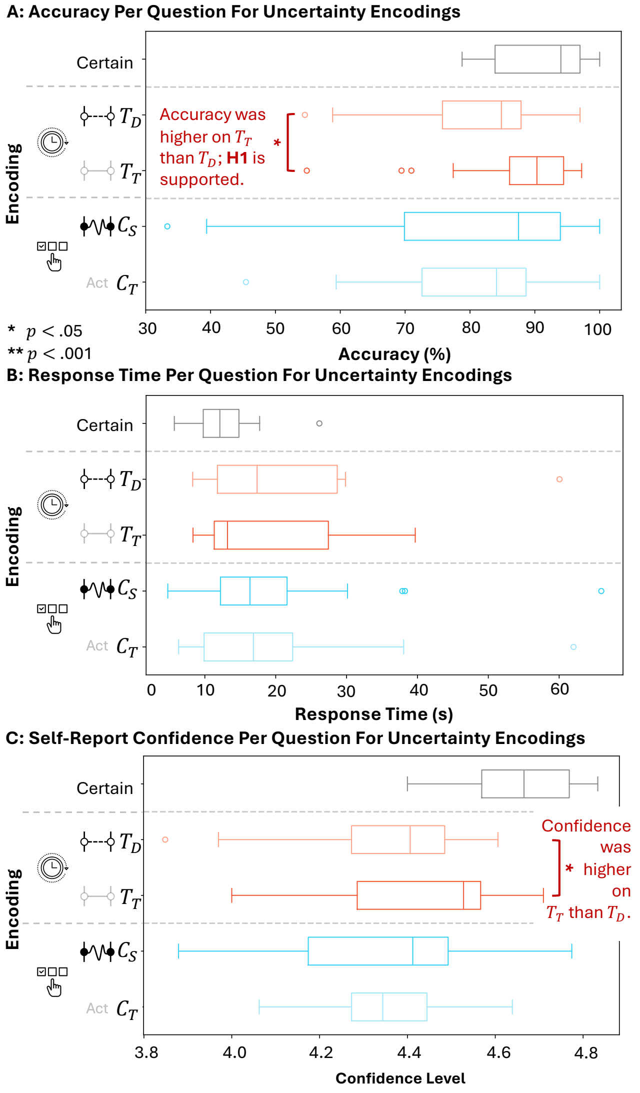

Figure 3: Plots for the accuracy (A), response time (B), and self-report confidence (C) on each encoding. Encodings are shorted for brevity.

RQ1: Temporal Uncertainty Encodings Participants were more accurate ($t_{(1096.2)} = 2.69$, $p = .004$, 95% CI [0.02, ∞]) and confident ($t_{(1109.5)} = 2.20$, $p = .01$, 95% CI [0.03, ∞]) on $T_{Transparent}$ than T_{Dashed} (we found no evidence they respond faster); **H1** is supported. ***In-person*** participants' preferences were split between temporal encodings: seven (39%) preferred $T_{Transparent}$, eight (44%) preferred T_{Dashed}, and three (17%) had no preference. Three participants (22%) preferred $T_{Transparent}$ because it was intuitive: "*[the time frame] is faded. It's like there and not there*," [P2] while only one (8%) preferred T_{Dashed} because it was intuitive, "*it's a little bolder and [...], well, that's obviously a gap*" [P6]. Although more participants preferred T_{Dashed}, encoding preference did not always align with performance. Twelve (92%) performed better on $T_{Transparent}$ or as good on $T_{Transparent}$ as T_{Dashed}. Only one participant (8%)—who, interestingly, preferred $T_{Transparent}$—performed best on T_{Dashed}.

For temporal uncertainty, $T_{Transparent}$ was *more effective* than T_{Dashed}.

RQ2: Categorical Uncertainty Encodings Quantitatively, we found no evidence that $C_{Transparent}$ was better than $C_{Sketchy}$ in terms of accuracy, confidence, or response time; **H2** is not supported. Additionally, ***in-person*** participants were split between the encodings: nine (50%) preferred

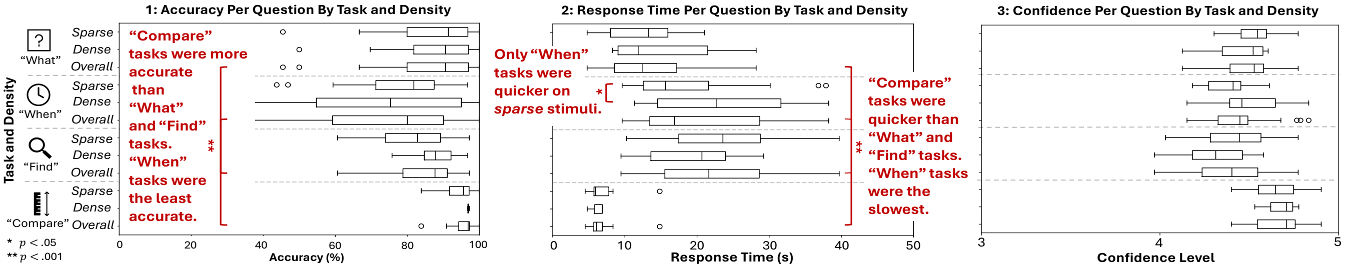

Figure 4: Plots of the accuracy (1), response time (2), and self-report confidence (3) for timeline evaluation tasks across stimuli densities.

$C_{Sketchy}$, eight (44%) preferred $C_{Transparent}$, and one (6%) had no preference. Eight participants (44%) thought $C_{Sketchy}$ was easier to see: "*[it is] harder to ignore*" [P13], but none (0%) thought it was intuitive. In comparison, five (63%) who preferred $C_{Transparent}$ believed it was intuitive: "*[the grayed out label] is like, the activity is a little bit unsure*" [P5].

$C_{Transparent}$ was a more *intuitive* categorical uncertainty encoding than $C_{Sketchy}$, but further evaluation is needed to determine effectiveness.

Because participation modality may have affected the statistical analysis of **RQ1** and **RQ2** (Fig. 3), we performed post-hoc two-tailed t-tests to investigate accuracy, confidence, and response time differences by uncertainty encoding between ***in-person*** and ***remote*** participants. Accuracy using $C_{Sketchy}$ ($t_{(8.96)} = 3.24$, $p = 0.01$, 95% CI [3.48, 19.53]) is the only statistically significant factor and can perhaps be attributed to the demographic differences between the two groups.

RQ3: Timeline Evaluation Tasks and Density in ARTs Although *sparse* stimuli contain fewer activities, only "When" tasks were completed quicker on *sparse* stimuli; **H3** is not supported. Additionally, we found "Compare" tasks, and not "What" tasks, had the highest accuracy regardless of ART density or uncertainty encoding (we excluded all but one "Compare" question from our analysis); **H4** is not supported. These quantitative results (summarized in Fig. 4) can possibly be attributed to our data exclusion protocols, lack of partial credit, and biases discovered during the qualitative analysis (***handling missingness*** and ***interpretation of linguistic uncertainty***).

RQ4: Perception of and Biases in ARTs After merging similar codes, we categorized the 10 unique codes into four themes, each of which likely altered ART interpretations and **online survey** responses.
1) Preconceived Notions of Human Behavior: All 18 in-person participants (100%) rationalized ARTs using their mental model of human behavior. All participants contextualized activities/durations: "*Some people might take half an hour because tea in India is different than tea [in America]*" [P5] or "*this person [...] brushes their teeth while showering*" [P12], frequently incorporating their existing cultural biases.
2) Interpreting Another's Data: Eleven participants (61%) reacted to the stimuli, questioning the validity of the data, "*It's not feasible that you would be brushing teeth for 30 minutes*" [P13], or the certainty of the recaller, "*You don't know how sure they were about their [activities]*" [P14]. Seemingly, uncertainty in the ART was partially interpreted as uncertainty in the data source: *"[It'd be easier] if you're reflecting on your data"* [P12].
3) Handling Missingness: Ten (56%) participants believed anything could have happened during times when no *certain* activity was recalled: "*He might be reading, or might be using his computer, or maybe doing something else*" [P16]. Additionally, six (33%) of these participants believed temporally uncertain activities could expand to fill times when no certain activity was recalled: "*there's no fixed time of the reading activity, so it could expand till 7:00. Or, if we stretch it out, it could start at 5:00*" [P4].
4) Interpretation of Linguistic Uncertainty: Some participants expressed confusion over the deliberately ambiguous framing of specific tasks (e.g., "When *could* this individual have...") to encourage inclusion of uncertain activities: "*does [this activity] fall under 'could have been' or 'was' using a computer?*" [P14]. This may have affected their accuracy because only responses selecting *all* applicable activities were correct. Based on the demographics of these participants, differences in ***handling missingness*** and ***linguistic uncertainty*** may stem from cultural or suspected primary language differences, but future work is warranted.

5 Discussion

Design Implications for ARTs: Because participants were inclined to ***handled missingness*** differently, ARTs may be improved by encoding missing data potentially preventing users of ARTs from making assumptions or distrusting the recaller. Additionally, because participants found it challenging to interpret ***another's data***, encoding the level of certainty of the recaller (e.g., altering the level of transparency) or using ARTs to encode self-recalls may increase confidence in the data.
Design Implications for Uncertainty on Timelines: We found ambiguation, $T_{Transparent}$, was an intuitive and effective temporal uncertainty encoding, in line with prior work [14]; additionally, we showed encoding temporal uncertainty using ambiguation was effective even with the addition of categorical uncertainty in ARTs. Thus, designers should consider using encodings similar to $T_{Transparent}$. Although five in-person participants believed $C_{Transparent}$ was an *intuitive* categorical uncertainty encoding, we found no evidence it was more effective than $C_{Sketchy}$. This could be attributed to the between-subjects design or unfamiliarity with categorical uncertainty because current timelines do not, but perhaps should, allow multiple event sequences represented in parallel, i.e., events scheduled concurrently are not represented differently than those not.
Limitations & Future Work: To prevent fatigue, we designed our study to not exceed 45 minutes which limited the evaluated factors. Additionally, our metrics for effectiveness provide opportunities for future work on study validation methodologies in timeline visualization because they did not account for partially correct answers or the accuracy and response time trade-off. Lastly, we used ARTs for the applications and impact they provide individuals and health researchers for tracking activity/behavioral patterns over time, however, understanding of recall data, as well as cultural differences, may have altered participant responses. Thus, while we demonstrated the influence of preconceived biases when interpreting timelines, this data may limit the breadth of the validity of our results, paving the way for future studies.

Hosting workshops for participants to encode their own data in ARTs may yield information on the intuitiveness of various uncertainty encodings while removing the biases associated with using another's data. Additionally, these encodings should be evaluated using less biased timelines, like those used for project planning or historical events. We hope future researchers are inspired to explore additional uncertainty encodings in timelines capturing fluctuating durations and multiple possible event sequences.

6 Conclusion

The results of our evaluation study indicate 1) transparency may be a good choice for conveying temporal uncertainty even in timelines with multiple kinds of uncertainty and 2) future work is needed to make conclusions about categorical uncertainty encodings in timeline visualizations. Our thematic analysis revealed that biases such as existing mental models and assumptions about the underlying data may impact how users view timelines, particularly those with uncertainty. This study lays the groundwork for encoding and evaluating multiple kinds of uncertainty in timelines, hopefully yielding more expressive and informative visualizations.

Acknowledgements

We thank Dr. Laura South, our colleagues, and reviewers for their advice and guidance. This research was supported in part by the Khoury College of Computer Sciences, Northeastern University and funded in part by the National Cancer Institute of the National Institutes of Health, grant number R01CA252966. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] ATUS home — bls.gov. https://www.bls.gov/tus/database.htm. [Accessed 10-04-2025].

[2] PAAWS Study — paawsstudy.org. https://www.paawsstudy.org/. [Accessed 20-04-2025].

[3] W. Aigner, S. Miksch, B. Thurnher, and S. Biffl. Planninglines: novel glyphs for representing temporal uncertainties and their evaluation. In *Ninth International Conference on Information Visualisation (IV'05)*, pp. 457–463, 2005. doi: 10.1109/IV.2005.97

[4] B. A. Aseniero, C. Perin, W. Willett, A. Tang, and S. Carpendale. Activity river: Visualizing planned and logged personal activities for reflection. In *Proceedings of the 2020 International Conference on Advanced Visual Interfaces*, AVI '20. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3399715.3399921

[5] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 2018. doi: 10.1111/j.2517-6161.1995.tb02031.x

[6] G.-P. Bonneau, H.-C. Hege, C. R. Johnson, M. M. Oliveira, K. Potter, P. Rheingans, and T. Schultz. *Overview and State-of-the-Art of Uncertainty Visualization*, p. 3–27. Mathematics and Visualization, 2014. doi: 10.1007/978-1-4471-6497-5$_1$

[7] N. Boukhelifa, A. Bezerianos, T. Isenberg, and J.-D. Fekete. Evaluating sketchiness as a visual variable for the depiction of qualitative uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2769–2778, 2012. doi: 10.1109/TVCG.2012.220

[8] M. Brehmer, B. Lee, B. Bach, N. H. Riche, and T. Munzner. Timelines revisited: A design space and considerations for expressive storytelling. *IEEE Transactions on Visualization and Computer Graphics*, 23(9):2151–2164, 2017. doi: 10.1109/TVCG.2016.2614803

[9] S. C. Castro, P. S. Quinan, H. Hosseinpour, and L. Padilla. Examining effort in 1D uncertainty communication using individual differences in working memory and NASA-TLX. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):411–421, 2022. doi: 10.1109/TVCG.2021.3114803

[10] A. Cimatti, A. Micheli, and M. Roveri. Timelines with temporal uncertainty. *Proceedings of the AAAI Conference on Artificial Intelligence*, 27(1):195–201, 2013. doi: 10.1609/aaai.v27i1.8601

[11] S. Di Bartolomeo, A. Pandey, A. Leventidis, D. Saffo, U. H. Syeda, E. Carstensdottir, M. Seif El-Nasr, M. A. Borkin, and C. Dunne. Evaluating the effect of timeline shape on visualization task performance. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, p. 1–12. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3313831.3376237

[12] M. Fernandes, L. Walls, S. Munson, J. Hullman, and M. Kay. Uncertainty displays using quantile dotplots or CDFs improve transit decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, p. 1–12. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3173574.3173718

[13] J. Fulda, M. Brehmer, and T. Munzner. Timelinecurator: Interactive authoring of visual timelines from unstructured text. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):300–309, 2016. doi: 10.1109/tvcg.2015.2467531

[14] T. Gschwandtner, M. Bogl, P. Federico, and S. Miksch. Visual encodings of temporal uncertainty: A comparative user study. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):539–548, 2016. doi: 10.1109/tvcg.2015.2467752

[15] S. Intille, C. Haynes, D. Maniar, A. Ponnada, and J. Manjourides. μEMA: Microinteraction-based ecological momentary assessment (EMA) using a smartwatch. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, p. 1124–1128. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2971648.2971717

[16] R. Kosara and S. Miksch. Metaphors of movement: A visualization and user interface for time-oriented, skeletal plans. *Artificial Intelligence in Medicine*, 22(2):111–131, 2001. doi: https://doi.org/10.1016/S0933-3657(00)00103-2

[17] J. Ludbrook. Should we use one-sided or two-sided p values in tests of significance? *Clinical and experimental pharmacology and physiology.*, 40(6), 2013-06. doi: 10.1111/1440-1681.12086

[18] C. E. Matthews, S. Kozey Keadle, S. C. Moore, D. S. Schoeller, R. J. Carroll, R. P. Troiano, and J. N. Sampson. Measurement of active and sedentary behavior in context of large epidemiologic studies. *Medicine Science in Sports Exercise*, 50(2):266–276, 2018. doi: 10.1249/mss.0000000000001428

[19] N. M. Moacdieh and N. Sarter. The effects of data density, display organization, and stress on search performance: An eye tracking study of clutter. *IEEE Transactions on Human-Machine Systems*, 47(6):886–895, 2017. doi: 10.1109/THMS.2017.2717899

[20] P. H. Nguyen, K. Xu, R. Walker, and B. W. Wong. Schemaline: Timeline visualization for sensemaking. In *2014 18th International Conference on Information Visualisation*, pp. 225–233, 2014. doi: 10.1109/IV.2014.14

[21] L. Padilla, M. Kay, and J. Hullman. Uncertainty visualization. *Wiley StatsRef: Statistics Reference Online*, pp. 1–18, 2021. doi: 10.1002/9781118445112.stat08296

[22] L. M. K. Padilla, S. H. Creem-Regehr, and W. Thompson. The powerful influence of marks: Visual and knowledge-driven processing in hurricane track displays. *Journal of Experimental Psychology: Applied*, 26(1):1–15, 2020. doi: 10.1037/xap0000245

[23] M. Rabbi, K. Li, H. Y. Yan, K. Hall, P. Klasnja, and S. Murphy. Revibe: A context-assisted evening recall approach to improve self-report adherence. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 3(4):Article 149, 2020. doi: 10.1145/3369806

[24] T. Sekino. Representation and comparison of uncertain temporal data based on duration. In *2018 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC)*, pp. 1–6, 2018. doi: 10.23919/PNC.2018.8579465

[25] M. Sharmin, A. Raij, D. Epstien, I. Nahum-Shani, J. G. Beck, S. Vhaduri, K. Preston, and S. Kumar. Visualization of time-series sensor data to inform the design of just-in-time adaptive stress interventions. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, p. 505–516. Association for Computing Machinery, New York, NY, USA, 2015. doi: 10.1145/2750858.2807537

[26] J. Smyth and A. Stone. Ecological momentary assessment research in behavioral medicine. *Journal of Happiness Studies*, 4:35–52, 2003. doi: 10.1023/A:1023657221954

[27] Y. Vaizman, K. Ellis, G. Lanckriet, and N. Weibel. Extrasensory app: Data collection in-the-wild with rich user interface to self-report behavior. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, p. 1–12. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3173574.3174128

[28] C. Ware. *Visual thinking for design*. Elsevier, 2010.

[29] K. Xu, S. Salisu, P. H. Nguyen, R. Walker, B. L. W. Wong, A. Wagstaff, G. Phillips, and M. Biggs. Timesets: Temporal sensemaking in intelligence analysis. *IEEE Computer Graphics and Applications*, 40(3):83–93, 2020. doi: 10.1109/mcg.2020.2981855