HaTS: Large-scale In-product Measurement of User Attitudes & Experiences with Happiness Tracking Surveys

Hendrik Müller
Google Australia Pty Ltd.
Level 5, 48 Pirrama Road
Pyrmont, NSW 2009, Australia
hendrik82@gmail.com

Aaron Sedley
Google Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043, USA
asedley@gmail.com

ABSTRACT

With the rise of Web-based applications, it is both important and feasible for human-computer interaction practitioners to measure a product's user experience. While quantifying user attitudes at a small scale has been heavily studied, in this industry case study, we detail best Happiness Tracking Surveys (HaTS) for collecting attitudinal data at a large scale directly in the product and over time. This method was developed at Google to track attitudes and open-ended feedback over time, and to characterize products' user bases. This case study of HaTS goes beyond the design of the questionnaire to also suggest best practices for appropriate sampling, invitation techniques, and its data analysis. HaTS has been deployed successfully across dozens of Google's products to measure progress towards product goals and to inform product decisions; its sensitivity to product changes has been demonstrated widely. We are confident that teams in other organizations will be able to embrace HaTS as well, and, if necessary, adapt it for their unique needs.

Keywords

Surveys; metrics; tracking; attitudes; large scale; HaTS.

Categories and Subject Descriptors

H.5.2. [**Information Interfaces and Presentation (e.g. HCI)**]: User Interfaces—benchmarking, evaluation/ methodology, standardization

1. INTRODUCTION

Human-computer interaction (HCI) practitioners employ a variety of research methods in their work, including evaluative research, exploratory research, behavioral analysis, and attitude measurement, among others. While measuring attitudinal data at a small scale for a given design or product (i.e., in the lab or field) has been studied heavily and is widely adopted in the HCI community, there have been fewer contributions towards a model to reliably track a product's attitudes over time and at a large scale.

In this industry case study, we are introducing a particular survey method, referred to as Happiness Tracking Surveys (HaTS), that is designed for ongoing tracking of user attitudes and experiences within the context of real-world product usage at a large scale. HaTS represents an optimized approach to data collection, sampling, and analysis. The short questionnaire instrument measures users' happiness with a given product, as part of gauging the overall quality of the user experience [23]. We refer to happiness as a set of metrics such as overall satisfaction, likelihood to recommend, perceived frustrations, and attitudes towards common product attributes, among others. Its design is grounded in an extensive body of questionnaire design research, substantial experimental testing, and best practices in data analysis, all with the goal to optimize validity, reliability, and sensitivity. Additionally, HaTS' random sampling approach ensures that collected metrics can be compared over time and across distinct user groups.

As HaTS is currently successfully deployed for ongoing user attitude tracking across dozens of Google products covering a wide range of categories (both consumer and enterprise products) and while being used by millions of users each, we are confident that user experience teams in other organizations will be able to embrace HaTS as well, and, if necessary, adapt it for their unique needs. However, note that our intention is not to question or replace existing user research practices, but rather complement those through the use of the presented large-scale attitudinal tracking method.

The remainder of this industry case study highlights relevant and related work, introduces the HaTS questionnaire instrument as well as its sampling approach, and then demonstrates its success by highlighting several applications of HaTS within Google.

2. RELATED WORK

In recent years, numerous questionnaires have been developed and continuously refined to measure a product's perceived quality on several dimensions, such as efficiency, effectiveness, helpfulness, learnability, satisfaction, and visual aesthetics. Some of the most commonly used ones (in order of their first publication) include the Computer User Satisfaction Inventory (CUSI) [8], the Questionnaire for User Interface Satisfaction (QUIS) [2], the After Scenario Questionnaire (ASQ) [19], the Software Usability Measurement Inventory (SUMI) [7], the Computer System Usability Questionnaires (CSUQ) [20], the System Usability Scale (SUS)

OZCHI '14, December 2–5, 2014, Sydney, NSW, Australia

http://dx.doi.org/10.1145/2686612.2686656

[1], and AttrakDiff [6]. These questionnaires were developed primarily with the intention of being used as part of usability evaluations or other small-scale product assessments. However, to our knowledge, they have rarely been used for ongoing attitude tracking in the context of real-world product usage at a large scale.

Nevertheless, methods exist to engage with a product's users in the context of actual product usage; one such commonly used method are feedback forms, which today are included across numerous Websites and Web applications. Most feedback forms are limited to one or a few open-ended questions, while some also include rating questions attempting to measure user's attitudes. However, the passive deployment of feedback forms to all users all the time suffers from bias towards those respondents that are experiencing a particular problem at any given time, i.e., respondents are not randomly sampled and the data being collected is likely less representative of the entire user base.

To measure user experiences at a large scale, some advances have recently been made, especially with regards to behavioral analysis through the use of log data. HEART was introduced as a framework for defining user-driven metrics regarding user happiness, engagement, adoption, retention, and task success [23]. While this work set the foundation for how to include large-scale measurement and analysis of user experiences into the product development process, the method for measuring the rather abstract construct of happiness was only explored briefly. In the context of HaTS, similar to HEART, we use the term happiness to describe a set of metrics that are attitudinal in nature and capture users' reactions to the entirety or parts of a product. Such metrics include overall satisfaction, likelihood to recommend, visual appeal, ease of use, and perceived speed, among others, which can be measured through survey-based methodologies.

While it initially appears easy and inexpensive to conduct surveys, overlooking key considerations in questionnaire design and the survey research process can yield skewed, biased, or entirely invalid survey results. Fortunately, insights related to the design of valid and reliable questionnaires has evolved significantly over the recent years due to the extensive work by social scientists. This includes a significant research body on different types of survey biases such as satisficing [9, 14, 10], acquiescence [28, 24], social desirability [26], and order effects [17, 30]. Particular attention has been paid to the design of appropriate question response options and rating scales [13, 5], and, more recently, to the visual design of surveys [3]. These survey research advances have recently been described in the context of conducting surveys in the field of HCI [21].

The HaTS survey method presented in this industry case study relies on these latest advances in questionnaire design and attempts to fill the gap of measuring attitudes at a large scale, in the context of real-world product usage, and over time through random sampling and the use of proactive survey invitations to aim for valid, reliable, and actionable data.

3. HATS OBJECTIVES

As Google's range of products expanded and teams increasingly focused on continually improving the user experience with these products, there was a desire to benchmark user happiness and to track teams' successes in improving user experiences through iterative development. To meet this primary need, we began developing HaTS in 2006. Since then, it has gone through several iterations and is now established as a standard across the entire company. The objectives of HaTS today can be summarized as follows, in order of importance:

1. To track changes in users' attitudes and perceptions over time (often weekly) and to associate those shifts to changes in the product or user base being surveyed. As such, this data identifies successes and failures of a product change, and helps pinpoint areas that require further exploration, focus, and improvement.

2. To collect open-ended feedback (both frustrations and areas of appreciation) from a representative sample of the product's user base. These data are then used to generate prioritized lists of frustrations as well as areas of appreciation to inform product strategy.

3. To characterize the product's user base, as well as to identify how users with different characteristics compare on those metrics tracked for the above goals.

4. To enable additional survey research with a representative sample of the entire user base in an extremely fast manner, i.e., through ad-hoc survey questions inserted for a short period of time (not to track longitudinally).

4. HATS SAMPLING AND INVITATION

The essence of surveying is sampling, i.e., to gather information about a population by obtaining data from a subset of that population. Depending on the population and number of respondents, estimates can be made at a certain level of precision and confidence. HaTS uses a probability sampling approach, the gold standard for achieving representative results, as users are randomly selected for survey invitation. The approach used for HaTS also borrows from the experience sampling method widely used in HCI [18].

Each week, a representative set of a product's users are randomly selected to be invited to take part in HaTS. To achieve this, the entire user base is randomly divided into distinct buckets for each week of the year. Randomization is based on individual users, instead of page or product views, to reduce the bias towards users that visit the same product frequently during the same week and instead give those that visit that product only once during the selected week a similar chance at being invited to the survey. To avoid effects of survey fatigue, the same user will not be invited to HaTS again for another 12 weeks after the survey invitation was previously exposed. As a result, during any given week, a maximum of about 8% of the entire user base may be invited to the survey.

The target sample size, i.e., the number of survey completions received, depends on the level of precision needed for reliable estimates and comparisons. For HaTS, as a best practice, we often aim for about 400 or 1000 responses for the time period of interest, which translates to approximately plus or minus 5% and 3% margins of error, respectively, with 95% confidence.

The survey mode for HaTS has exclusively been Web-based. To invite potential respondents, HaTS is exposed either through a link, a banner, or a mole in the product

to users being sampled. The invitation is visually differentiated from other content on the page, and positioned so that it is easily visible when the page loads, to ensure potential respondents are actively reached out to. However, HaTS avoids model pop-up invitations that require users to respond to the pop-up before they can continue to the rest of the site, as that interrupts users' normal workflows and often upsets potential respondents. The language for the invitation is set to "Help us improve [product]" with a "Take our survey!" call to action (and equivalent translations for other languages); this neutral wording encourages any user to respond to the survey, not just those interested in venting or praising. As such, the invitation text also highlights why it is important for the user to take this survey, as it is helping to improve the product for users just like them. One of the fundamental strengths of HaTS remains its attempts to enhance data validity by inviting people as they are using the product itself, therefore the responses directly reflect users' attitudes and perceptions in context of their experiences with the actual product. In contrast, surveys completed well after the experience being studied may suffer from imperfect recall, retrospective bias, and intermediary experiences that affect responses.

5. HATS QUESTIONNAIRE ELEMENTS

The design of the HaTS questionnaire instrument (see Figure 10 for the complete questionnaire) follows established guidelines to optimize the reliability and validity of responses [21]. It minimizes common biases such as satisficing [9, 14, 10], acquiescence [28, 24], social desirability [26], and order effects [17, 30]. It also relies on an extensive body of research regarding the design of scales and responses options [13]. The visual design of the HaTS questionnaire is optimized for increased usability, however, without the use of unnecessary images or other themes, to avoid introducing additional biases as identified by Couper [3].

To avoid question order biases (i.e., questions earlier in the survey to influence questions later in the survey [17]), the HaTS questionnaire follows a "funnel" approach from broad and high-level to more specific and personal questions. In the beginning of the questionnaire, we include questions directly related to the survey topic and ask about attitudes and feedback about product as a whole (to avoid potential biases resulting from questions that ask about specific aspects of the product). These initial questions are also important to help build rapport with the respondent. After high-level aspects have been assessed, the questionnaire then dives into common product attributes as well as product-specific tasks. Finally, questions about respondents' characteristics are asked about towards the end as they may be perceived as more sensitive by some. Furthermore, pagination is used as a tool to group related questions and to reduce context switching for the respondents. HaTS' first page assesses the respondent's overall attitudes and experiences with the product, the second page explores common attributes and product-specific tasks, the third page asks about respondent characteristics, with the fourth and last page reserved for ad-hoc questions. In the remainder of this section we discuss each of the HaTS questions, in the order they appear in the questionnaire.

At the beginning of HaTS, the purpose and importance of the questionnaire is explained using a few sentences, such as: "Thank you for offering your feedback on [product]. Understanding your experiences and opinions helps [product] make this product better for you and other users." Additionally, information regarding our company's privacy policy is presented and it is often mentioned how much time it usually takes to complete the questionnaire.

5.1 Overall Product Satisfaction and Likelihood to Recommend

The core of HaTS lies in measuring users' overall attitudes with a given product over time, which is achieved by tracking the metric of satisfaction. Satisfaction is measured through the question "Overall, how satisfied or dissatisfied are you with [product]?" (see Figure 1). Note that the question text refers to satisfaction in a neutral way, through calling out both "satisfied" and "dissatisfied." As the construct of satisfaction is bipolar in nature (i.e., it naturally has two opposite extremes and a neutral midpoint), a 7-point scale is used to optimize validity and reliability, while minimizing the respondents' efforts [13, 16]. The scale is fully labeled without the use of any numbers to ensure respondents entirely focus on the meaning of the answer options [15]. Scale items are displayed horizontally to minimize order biases and are equally spaced [29], both semantically as well as visually [3]: "Extremely dissatisfied," "Moderately dissatisfied," "Slightly dissatisfied," "Neither satisfied nor dissatisfied," "Slightly satisfied," "Moderately satisfied," and "Extremely satisfied." Note the user of "Neither satisfied nor dissatisfied" instead of "Neutral" as the midpoint, to minimize the effect of satisficing [9]. Furthermore, the negative extreme is listed first to allow for a more natural mapping to how respondents interpret bipolar constructs [30]. Even though each of these best practices are derived from the extensive body of literature, we replicated several of these experiments with our own HaTS questions in the contexts the survey is being used.

The satisfaction question is the only question within HaTS that is required to be answered (i.e., the respondent cannot proceed with the survey without answering this question), as without answering this question the primary goal of HaTS cannot be met.

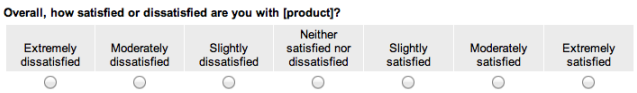

Figure 1: Overall satisfaction measured on a fully-labeld 7-point scale.

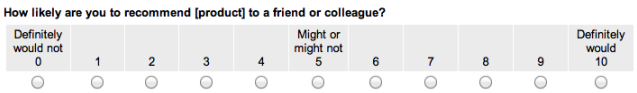

Figure 2: Overall likelihood to recommend measured on the typical 11-point scale.

Even though we believe that the construct of satisfaction provides sufficient insights into the overall attitude towards a product, in some cases, HaTS also includes a question that measures the respondent's likelihood to recommend the product to others: "How likely are you to recommend [product] to a friend or colleague?" (see Figure 2). In accordance with the widely used net promoter question format [22], a

partially-labeled 11-point scale is then used with "Definitely would not" and "Definitely would" as labels on the extreme points and "Might or might not" as the midpoint. However, as there is considerable skepticism about the reliability of the net promoter question due to its design and analysis approach (e.g., [4]), and as it is unclear if likelihood to recommend provides a metric that is distinct from satisfaction, this question is only included on an as-needed basis (e.g., when a divergence between satisfaction and likelihood to recommend is expected).

5.2 Open-ended Frustrations and Areas of Appreciation

To gather qualitative data about users' experiences with a given product, HaTS also includes two open-ended questions. Respondents are asked to describe the most frustrating aspects of the product and new capabilities they would like to see ("What, if anything, do you find frustrating or unappealing about [product]? What new capabilities would you like to see for [product]?," see Figure 3), as well as what they like the most about the product ("What do you like best about [product]?," see Figure 4). Note that it may appear that the frustrations questions is double-barrelled in its current design. However, through several experiments across several products, we determined that asking about experienced frustrations and needed new capabilities in the same question increased the response quantity (i.e., a higher percentage of respondents provided input) and quality (i.e., responses were more thoughtful in their description and contained further details) and minimized the analysis effort as compared to using two separate questions. Further experiments helped us determine the specific words to be used when asking about problems and missing functionality, leading to the use of "frustrating or unappealing" and "new capabilities."

To avoid question order biases, these open-ended questions are asked directly after the overall satisfaction question at the beginning of HaTS. As identifying product opportunities is one of the main goals for HaTS, respondents are encouraged to spend more effort on the frustrations questions. This is achieved by listing the frustrations before the areas of appreciation questions and by increasing the size of the frustrations answer text box (as the size of the text box suggests the approximate length of the expected response [3]). Even though these questions are not compulsory, HaTS calls them out as "(Optional)" in the beginning of the question. Evaluated again through several experiments replicated across several products, this ensures that, first, respondents do not drop off the survey if they perceive an open-ended response as too much effort, and, second, to minimize random responses (e.g., "asdf") that slow down the analysis process. Note that, even though expected, the addition of the "(Optional)" label did not result in significantly less responses to that question, while response quality increased. We have consistently received question response rates between 40 and 60% for these open-ended questions when used in this format.

5.3 Satisfaction with Common Attributes and Product-specific Tasks

In addition to measuring overall attitudes, HaTS also assesses different components of the user experience, in particular, satisfaction with attributes that are common to most

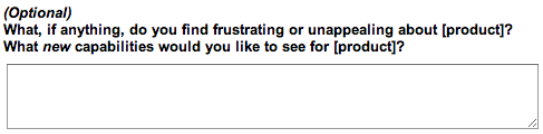

Figure 3: Open-ended question about frustrations and new capabilities to be added to the product.

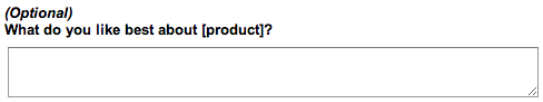

Figure 4: Open-ended question capturing areas of appreciation.

products as well as satisfaction with product-specific tasks. Attributes such as perceived "ease of use," "technical reliability," "features & capabilities," "visual appeal," and "speed" are measured through a similar question as the overall satisfaction question ("How satisfied or dissatisfied are you with [product]? in the following areas"); however, a grid is used with each of the attributes listed as rows and the usual 7-point scale as columns (see Figure 5). Note that no "don't know" or other opt-out option is provided, as all of those attributes apply to each product being measured, and, hence, the respondents should have a valid attitude towards them. Not including an opt-out option again reduces the effect of satisficing [11, 25]. However, if some of the attributes do not apply to a given product, they are simply excluded.

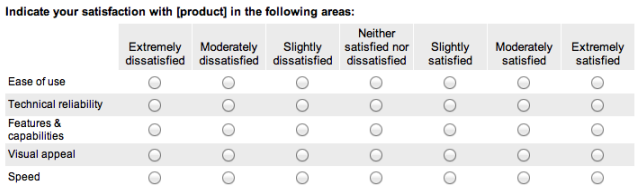

Figure 5: Satisfaction with common product attributes.

Second, HaTS asks about the level of satisfaction for a set of common tasks the product being measured attempts to support ("How satisfied or dissatisfied are you with doing the following tasks in [product]?," see Figure 7). Tasks are intended to be things that the measured products or other products similar to it would naturally be able to address; however, tasks should not refer to specific features. Example tasks may be "Previewing any kind of file" or "Using tables in a document." As not all respondents may have previously attempted all of the listed tasks (and hence would not have reliable attitudes towards them), on the page prior, HaTS first asks respondents to select those tasks they have attempted over the last month ("In the last month, which of the following tasks have you tried to accomplish with [product]?," see Figure 6). Note that the use of this reference period ensures that their experience with that tasks is relatively recent to warrant reliable satisfaction scores. If a respondent checks more than five of the listed tasks, a maximum of five are randomly selected for their satisfaction assessment in the subsequent question. Keeping the num-

ber of rows in the grid question row minimizes the effect of satisficing, in particular straight-lining [12]. To further discourage satisficing in the form of straight-lining, alternate row shading is used. Note that for all three in this section described questions, the order of the attributes and tasks are randomized across respondents to avoid response order effects [30].

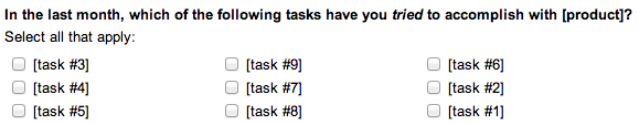

Figure 6: Typical product task selection question with answer options being randomized.

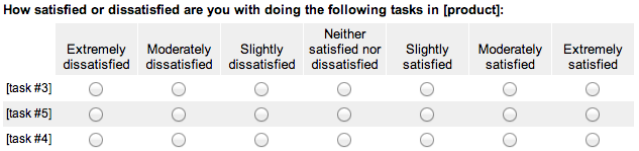

Figure 7: Satisfaction with those product-specific tasks selected in the question shown in Figure 6.

5.4 Respondent Characteristics

HaTS also asks questions for the respondent to self-report some of their characteristics in the context of using the product being measured. During the analysis, data from such questions can be used to compare attitudes across distinct user groups or to track changes in the user base over time. Questions often included may ask about first-time usage (e.g., "How many months ago did you start using [product]?" with a numeric open-ended text box, see Figure 8) and the frequency of usage (e.g., "In the last month, on about how many days have you used [product]?" with a numeric open-ended text box, see Figure 9). A numeric open-ended text box (i.e., a text field that is restricted to entering only numbers) is used to increase reliability, as compared to providing the user with a set of predefined answer options. Evaluated through experiments, we explicitly chose to ask about "in the last month" to refer to a specific time frame instead of asking the respondent to average when being asked about "typical" or "on average" usage.

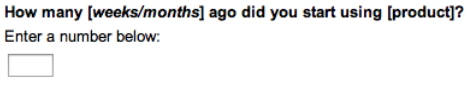

Figure 8: Respondent characteristics: Self-reported time since having started using the product.

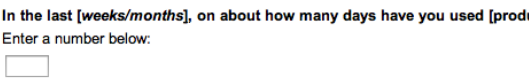

Figure 9: Respondent characteristics: Self-reported approximate usage frequency.

Additionally, questions that are specific to the domain of the product may be added (e.g., asking about the number of files stored with the product for an online file storage product, questions about the usage of related products, or questions about the savviness or experience with the domain). Instead of directly asking the respondent about such characteristics of their product usage, it is preferred to pipe data directly into the survey database. Common examples for piping include the product version the user is using and their language or country settings.

5.5 Ad-hoc Questions

Finally, HaTS is set up so that ad-hoc questions can easily be inserted for a short period where trend analysis is not needed. To ensure that such ad-hoc questions do not influence the standard HaTS questions and its paradata in any way, these ad-hoc questions are added on an additional page at the end of the survey, while the survey is set as completed already before that. This approach allows the researcher to explore specific aspects of the product as needed, while easily reaching a randomly selected sample of the product's user base. As a result, this has often been used in the requirements gathering stage.

6. USAGE AND APPLICATIONS OF HATS

HaTS has been used in a variety of ways at Google to aid product development and optimize users' experiences. One kind of successful application of HaTS has been to measure the short-term effect of UI changes on users' attitudes, a phenomenon known as "change aversion" [27]. As for example for the redesign launch of Google Drive, we compared average satisfaction levels in the weeks prior to and following an existing product's re-launch, showing the intensity, duration, and resolution of attitudinal changes due to users' initial experiences with the modified product. By identifying the degree to which users react negatively to the way changes are introduced, product teams can refine their future launch strategies, and even see whether certain changes are fundamentally degrading users' experiences beyond a natural adjustment period. Identifying and measuring change aversion across products has led to the development and adoption of a change management framework that minimizes users' pain and improves launches' prospects for success. After the change-influenced adjustment period, HaTS provides a new benchmark satisfaction level and tells us whether the improvements that were launched have made a meaningful and lasting positive attitudinal impact.

One of the strongest uses of HaTS is comparing experimental product versions to control versions. By randomly assigning users to different product versions and surveying each group as they use the product, we obtain clean comparisons of users' attitudes toward each version. Such attitudinal metrics can yield significant insight to A/B experiment evaluations, sometimes disambiguating an experiment's impact on user happiness where standard behavioral comparisons do not provide a reliable signal. Along with trend analysis and A/B comparisons, products use HaTS to compare differences in attitudes across geographies, by features, use cases, and user segments. For example, one Google product?s users in a particular language were significantly less satisfied than users in other languages, leading to an investigation that uncovered numerous linguistic and functional issues with the interface in that language.

Furthermore, open-ended questions in HaTS provide a clear, quantifiable view of users' actual experiences. These often represent the most insightful data collected through HaTS as it sheds some light into the "why" for certain attitudes being reported. Hundreds of users' responses are manually coded and then frequency-sorted, providing both a reliable prioritization of dissatisfaction causes and areas of appreciation, and qualitative insights that closed-ended questions do not yield. HaTS open-ended responses come from a more representative set of users than traditional "Send feedback" methods (which tend to be used more frequently by users experiencing a problem), thus providing a wider range of views to inform product decision-making. Insights from such an analysis has informed prioritization and product strategy in many cases, such as Google Drive.

HaTS has also been used to gauge users' awareness of features and capabilities. Another Google product's HaTS found that a majority of users were unaware of some of features measured, helping diagnose reasons for low usage and disambiguating low awareness from perceived utility and usage difficulty. Additionally, HaTS has been used to clarify the sometimes opaque nature of user behavior. One product's behavioral log data were linked to HaTS attitudinal data, yielding insights where users' self-reported satisfaction did not match traditional behavioral signals. These findings helped refine existing models used to classify specific behaviors as positive or negative signals.

Finally, HaTS also lets us measure relationships between variables, and which specific product dimensions account for the most variance in high-level attitudes. In another Google product's case involving a redesigned UI released on a trial basis to a subset of users, HaTS identified a strong correlation between perceived speed and users' overall satisfaction, convincing the team to further decrease latency before launching the redesign to all users. By co-analyzing attitudinal metrics with user characteristics, we can see which user groups are least satisfied with a product, such as users of a particular feature or novices vs power users.

7. CONCLUSION

For a variety of reasons, HaTS has proven to be a useful, high quality method for measuring, tracking and comparing users' attitudes at a large scale, and one that can be effectively adopted by others who endeavor to better understand users' attitudes and experiences. From the outset, HaTS has used probability sampling, the gold standard among survey researchers for achieving representative results for a given population. Users are sampled at the moment they are actually using the product, ensuring that their responses accurately reflect their true experiences, unaffected by memory bias.

HaTS' question order, question text, response scales, as well as several other questionnaire design details follow best practices based on extensive academic experimentation to optimize data validity and reliability. Thus, we reduce the potential for survey satisficing behavior and associated biases such as acquiescence, social desirability, respondent fatigue, and overall non-differentiation. We have conducted extensive cognitive pretesting with several HaTS questionnaires, identifying respondents' areas of confusion and misinterpretation, and refining the surveys accordingly. In addition, real-world 50/50 experiments have further validated the effect of various question and scale formats, to help us avoid approaches that skew data from their natural distributions.

As demonstrated in this case study, HaTS has successfully been used to measure change aversion throughout product updates, comparing attitudes towards different product versions, understanding top frustrations and areas of appreciation to inform product strategy, gauging users' awareness with aspects of a product, and measuring relationships between attitudes and user and usage variables. HaTS can often be used to initially identify high-level insights that can be followed by in-depth research through more in-depth (meaning smaller-sample) methods.

Although the theoretical underpinnings of HaTS are solid, and continual improvements have further strengthened the platform, several challenges remain that inform our future work. First, while we have demonstrated the success of HaTS through the insights it has provided for numerous products over the years, it has not formally been evaluated for its validity and reliability. To establish HaTS as a standardized questionnaire for large-scale in-product attitudinal tracking, this should be the next step.

Our standard "Take our survey!" link is not particularly prominent in some product's user interfaces, and may be easily ignored or not be seen at all by many users, thus raising the potential for non-response bias. We may experiment with more salient designs, including placing the initial question and response choices directly in products' pages, with a follow-up invitation to complete the remainder of the survey.

The analysis of large quantities of open-ended feedback is another challenge, in that it is a time-consuming manual process. It is worth splitting such efforts among multiple human coders, or even crowdsourcing such tasks via mechanical turk, provided that sufficient inter-coder reliability can be established. When we gather thousands of pieces of feedback, we generally randomly sub-sample a few hundred of these responses to provide a representative summary with an adequate precision level.

Even though HaTS is based on random sampling among active users, the self-selected nature of survey participation can result in non-response bias that skews results from the true underlying population values. In particular, more frequent visitors will see the survey invitation more often than infrequent visitors, and this increased exposure will likely over-represent heavy users.

Non-response investigations are important to accurately representing each product's entire user base, and when we discover user characteristics that differ between the sample data and the full population - and that correlate with variance in key survey metrics - it is valuable to post-stratify (weight) such characteristics to reflect their known population values. That said, the impact of non-response bias is lessened when comparing metrics over time, or across experimental conditions, where such bias is equally distributed in the comparison groups, and differences between groups can be attributed solely to time or product differences.

HaTS, as a means to evaluate and monitor user satisfaction and users' likes and dislikes is based on a strong foundation of both academic experimentation and established heuristics. The use of HaTS across products, and its proven sensitivity to identify meaningful attitudinal changes over time and between user groups and experimental conditions testify to the utility of the platform to inform decision making to improve users experiences as products iterate over

time. We believe other organizations can yield significant value by adopting HaTS, adjusting it to their specific needs, and continuing to refine the platform for high quality, actionable results.

8. ACKNOWLEDGMENTS

Thanks to Daniel Russell, Robin Jeffries, Kathy Baxter, Elizabeth Ferrall-Nunge, Marianne Berkovich, and others for their support and guidance with HaTS, and to the many other user experience researchers within Google for applying this instrument across the company.

9. REFERENCES

[1] J. Brooke. SUS: A quick and dirty usability scale. *Usability evaluation in industry*, 189:194, 1996.

[2] J. P. Chin, V. A. Diehl, and K. L. Norman. Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the SIGCHI*, CHI '88, pages 213–218. ACM, 1988.

[3] M. Couper. *Designing effective web surveys*. Cambridge University Press Cambridge, 2008.

[4] D. B. Grisaffe. Questions about the ultimate question: Conceptual considerations in evaluating reichheld's net promoter score (NPS). In *Journal of Consumer Satisfaction, Dissatisfaction, and Complaining Behavior: CS/D&CB*, pages 36–53, 2007.

[5] R. M. Groves, E. Singer, J. M. Lepkowski, S. G. Heeringa, and D. F. Alwin. *Survey methodology*. The University of Michigan Press, 2004.

[6] M. Hassenzahl, M. Burmester, and F. Koller. Attrakdiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In *Mensch & Computer 2003*, pages 187–196. Springer, 2003.

[7] J. Kirakowski and M. Corbett. SUMI: The software usability measurement inventory. *British journal of educational technology*, 24(3):210–212, 1993.

[8] J. Kirakowski and A. Dillion. The computer user satisfaction inventory. *Proceedings from the IEE: Evaluation Techniques for Interactive System Design*, 1987.

[9] J. A. Krosnick. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5:213–236, 1991.

[10] J. A. Krosnick. Survey research. *Annual review of psychology*, 50(1):537–567, 1999.

[11] J. A. Krosnick. The causes of no-opinion responses to attitude measures in surveys: They are rarely what they appear to be. *Survey nonresponse*, pages 87–100, 2002.

[12] J. A. Krosnick and D. F. Alwin. A test of the form-resistant correlation hypothesis ratings, rankings, and the measurement of values. *Public Opinion Quarterly*, 52(4):526–538, 1988.

[13] J. A. Krosnick and L. R. Fabrigar. Designing rating scales for effective measurement in surveys. *Survey measurement and process quality*, pages 141–164, 1997.

[14] J. A. Krosnick, S. Narayan, and W. R. Smith. Satisficing in surveys: Initial evidence. *New Directions for Evaluation*, 1996(70):29–44, 1996.

[15] J. A. Krosnick and S. Presser. Question and questionnaire design. *Handbook of Survey Research. 2nd edition. Bingley, UK: Emerald*, pages 263–314, 2010.

[16] J. A. Krosnick and A. M. Tahk. The optimal length of rating scales to maximize reliability and validity. *Unpublished manuscript, Stanford University*, 2008.

[17] E. L. Landon. Order bias, the ideal rating, and the semantic differential. *Journal of Marketing Research*, 8(3):375–378, 1971.

[18] R. Larson and M. Csikszentmihalyi. The experience sampling method. In H. T. Reis, editor, *Naturalistic Approaches to Studying Social Interaction*, volume 15 of *New Directions for Methodology of Social and Behavioral Science*, pages 41–56. Jossey-Bass, San Francisco, CA, USA, 1983.

[19] J. R. Lewis. Psychometric evaluation of an after-scenario questionnaire for computer usability studies: The ASQ. *SIGCHI Bulletin*, 23(1):78–81, 1991.

[20] J. R. Lewis. IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *Int. J. Hum.-Comput. Interact.*, 7(1):57–78, 1995.

[21] H. Müller, A. Sedley, and E. Ferrall-Nunge. Survey research in HCI. In J. Olson and W. Kellogg, editors, *Ways of Knowing in HCI*, pages 229–266. Springer, New York, NY, USA, 2014.

[22] F. F. Reichheld. The one number you need to grow. *Harvard Business Review*, 2003.

[23] K. Rodden, H. Hutchinson, and X. Fu. Measuring the user experience on a large scale: User-centered metrics for web applications. In *Proceedings of the SIGCHI*, CHI '10, pages 2395–2398. ACM, 2010.

[24] W. E. Saris, J. A. Krosnick, and E. M. Shaeffer. Comparing questions with agree/disagree response options to questions with construct-specific response options. *Unpublished manuscript, University of Amsterdam*, 2005.

[25] N. C. Schaeffer and S. Presser. The science of asking questions. *Annual Review of Sociology*, 29:65–88, 2003.

[26] B. R. Schlenker and M. F. Weigold. Goals and the self-identification process: Constructing desired identities. *Goal concepts in personality and social psychology*, pages 243–290, 1989.

[27] A. Sedley and H. Müller. Minimizing change aversion for the Google Drive launch. In *CHI '13 Extended Abstracts*, CHI EA '13, pages 2351–2354. ACM, 2013.

[28] D. H. Smith. Correcting for social desirability response sets in opinion-attitude survey research. *The Public Opinion Quarterly*, 31(1):87–94, 1967.

[29] R. Tourangeau. *Cognitive science and survey methods*, volume 73. National Academy Press Washington, 1984.

[30] R. Tourangeau, M. P. Couper, and F. Conrad. Spacing, position, and order interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly*, 68(3):368–393, 2004.

Thank you for offering your feedback on [product].
Understanding your experiences and opinions helps [company] make this product better for you and other users.

Overall, how satisfied or dissatisfied are you with [product]?

Extremely dissatisfied	Moderately dissatisfied	Slightly dissatisfied	Neither satisfied nor dissatisfied	Slightly satisfied	Moderately satisfied	Extremely satisfied
○	○	○	○	○	○	○

How likely are you to recommend [product] to a friend or colleague?

Definitely would not 0	1	2	3	4	Might or might not 5	6	7	8	9	Definitely would 10
○	○	○	○	○	○	○	○	○	○	○

(Optional)
What, if anything, do you find frustrating or unappealing about [product]?
What *new* capabilities would you like to see for [product]?

(Optional)
What do you like best about [product]?

How satisfied or dissatisfied are you with [product] in the following areas?

	Extremely dissatisfied	Moderately dissatisfied	Slightly dissatisfied	Neither satisfied nor dissatisfied	Slightly satisfied	Moderately satisfied	Extremely satisfied
Ease of use	○	○	○	○	○	○	○
Technical reliability	○	○	○	○	○	○	○
Features & capabilities	○	○	○	○	○	○	○
Visual appeal	○	○	○	○	○	○	○
Speed	○	○	○	○	○	○	○

In the last month, which of the following tasks have you *tried* to accomplish with [product]?
Select all that apply:

- ☐ [task #2]
- ☐ [task #1]
- ☐ [task #3]
- ☐ [task #6]
- ☐ [task #4]
- ☐ [task #5]

How satisfied or dissatisfied are you with doing the following tasks in [product]:

	Extremely dissatisfied	Moderately dissatisfied	Slightly dissatisfied	Neither satisfied nor dissatisfied	Slightly satisfied	Moderately satisfied	Extremely satisfied
[task #2]	○	○	○	○	○	○	○
[task #1]	○	○	○	○	○	○	○
[task #3]	○	○	○	○	○	○	○

How many [*weeks/months*] ago did you start using [product]?
Enter a number below:

In the last [*weeks/months*], on about how many days have you used [product]?
Enter a number below:

Figure 10: HaTS questions in their order of appearance.