“Who doesn’t like dinosaurs?” Finding and Eliciting Richer Preferences for Recommendation

Tobias Schnabel
toschnab@microsoft.com
Microsoft
Redmond, WA, USA

Gonzalo Ramos
goramos@microsoft.com
Microsoft
Redmond, WA, USA

Saleema Amershi
samershi@microsoft.com
Microsoft
Redmond, WA, USA

ABSTRACT

Real-world recommender systems often allow users to adjust the presented content through a variety of preference elicitation techniques such as “liking” or interest profiles. These elicitation techniques trade-off time and effort to users with the richness of the signal they provide to learning component driving the recommendations. In this paper, we explore this trade-off, seeking new ways for people to express their preferences with the goal of improving communication channels between users and the recommender system. Through a need-finding study, we observe the patterns in how people express their preferences during curation task, propose a taxonomy for organizing them, and point out research opportunities. We present a case study that illustrates how using this taxonomy to design an onboarding experience can lead to more accurate machine-learned recommendations while maintaining user satisfaction under low effort.

CCS CONCEPTS

• **Human-centered computing → Interaction paradigms**; • **Information systems** → *Personalization*; • **Computing methodologies** → *Machine learning*.

KEYWORDS

user preferences, onboarding experiences, user control

ACM Reference Format:
Tobias Schnabel, Gonzalo Ramos, and Saleema Amershi. 2020. “Who doesn’t like dinosaurs?” Finding and Eliciting Richer Preferences for Recommendation. In *Fourteenth ACM Conference on Recommender Systems (RecSys ’20), September 22–26, 2020, Virtual Event, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3383313.3412267

1 INTRODUCTION

Most content services such as e-commerce websites, online news providers or bookmarking services try to offer support in the form of personalized recommendations. An aspect of growing importance is the ability of people to exert control in these systems by expressing their preferences to them. Techniques for supporting people in expressing their preferences to these personalization services often trade-off user effort with the richness of information needed by machine learning components driving those personalization experiences.

On the lower end of the information richness spectrum are systems that support simple user feedback, such as binary (“liking” or “pinning”) or ordinal signals (e.g., 1-5 star ratings). The downside with this strategy is that these signals only indirectly serve as an indicator of a person’s interests and recent work has shown that people are more satisfied with services they feel they have more control over [39]. Moreover, these elicitation techniques have low information fidelity by design and hence are limited in their potential to improve machine-learned recommendations [18, 28, 31]. Some systems attempt to amplify these weak signals by enabling simple feedback on many dimensions, which can take the form of simple onboarding questionnaires (e.g., in streaming video providers such as Hulu) or settings (e.g., privacy controls in social networks such as Facebook or Twitter). While these techniques increase the signal from the user, they can also create a tedious and mentally demanding user experience and in most cases are entirely skipped or ignored [24]. On the upper end of the spectrum are systems such as chat bots that allow people to have full-blown conversations about their preferences. Although these natural language systems support rich interaction, they can pose a significant interaction cost to users [31] and are still at an early developmental stage [16].

We explore the space in between the ends of the information richness spectrum, from simple signals to high-bandwidth, semantically rich communication. Specifically, we aim to find ways to design preference elicitation instruments striking a balance between user effort and accurate recommendations. To this end, we present a need-finding study to understand how people articulate the reasoning behind their preferences on two curation tasks in the domains of images and text articles. We present a case study of *TellY*, an efficient preference elicitation instrument that is based on the taxonomy developed in the need-finding study. We conduct a crowd-sourced study comparing TellY to multiple traditional preference elicitation methods that trade-off user effort and signals differently. Our results show that TellY enables more accurate machine learning-based recommendation predictions while requiring comparable or lower levels of user effort to obtain.

2 RELATED WORK

Our work touches on two subareas of recommender systems – feedback signals that are used for training such systems and preference elicitation techniques for enabling people to express their preferences to the system.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

RecSys ’20, September 22–26, 2020, Virtual Event, Brazil

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-7583-2/20/09…$15.00
https://doi.org/10.1145/3383313.3412267

2.1 Feedback signals in recommender systems

Past research distinguishes between two main types of feedback signals for training recommender systems depending on how they are obtained [17]. *Implicit* feedback signals are collected passively by recording people's interactions with a system, such as clicking, hovering, or scrolling [15]. *Explicit* feedback signals are collected actively, for example by asking people to provide ratings or label items. The problem of merely relying on implicit feedback signals is the so-called *cold-start problem* where signals are not available when a new user joins [35]. Also, since these signals are collected passively, people have little control over what information they want to be used for their recommendations. In contrast, explicit signals require active user participation which increases effort for users. However, recent work has shown that users are willing and able to provide explicit feedback when it gives them a sense of control over the behaviors of their recommendations, which in turn results in an increase in user satisfaction [39]. In this paper, we focus on elicitation techniques for explicit signals, but note that any explicit signal can be used in conjunction with other implicit feedback signals as they become available.

One type of explicit signal that has been studied well as side information for training recommender systems is personality traits. For example, with their "Tune-A-Find" system, Ferwerda et al. [9] show how there is a correlation between a user's chosen taxonomy for a song, and a subset of the five-factor model (FFM) or "big five" personality dimensions: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism [25]. In other work, Hu and Pu [13] compared movie predictions stemming from recommender systems additionally using either item-based or personality-based feedback. While they found that prediction accuracy was similar, participants in the personality-based condition expressed they spent less effort, and were more likely to reuse the system. Hu and Pu [14] further look into the benefits of integrating both item scores and personality to improve collaborative filtering systems.

Regardless of what algorithm or signal type one chooses, recommender systems must consider privacy concerns [33]. In general, recommender systems have to trade-off the richness of information they elicit and employ to make accurate predictions with the potential risks to privacy – often being referred to as the *privacy-personalization trade-off* [2, 23]. Previous work has shown that people's decisions of how to balance these goals are influenced not only by their trust in the system [21], but also by the overall user experience with the system and the benefits it provides [5, 20]. As such, we also consider the perceived impact on user privacy in our case study.

2.2 Preference elicitation techniques

Preference elicitation techniques can be ordered with respect to their complexity. Starting with lower complexity techniques, category-based elicitation methods ask people to give select categories of interest from a fixed set of categories [27]. Because of their low complexity, these techniques are commonly used in practice to onboard new users. For example, new users on Hulu.com can choose from categories such as "Big Personalities", "Thrills & Chills", or "Edgy Animation". Item-based techniques ask people to provide ratings for a given or self-specified set of items. Their granularity can range from binary (e.g., thumbs up / down) over star ratings to continuous sliders. However, ratings are known to be inconsistent and have limited psychological backing [1, 29]. Although not as popular, item-based elicitation can also be cast as a comparative task where people specify their preferences either between pairs of items [7], or even between groups of items [11, 38]. Of medium complexity are personality-based elicitation techniques [9, 30] that profile people via common personality instruments, such as TIPI [10]. Of slightly higher complexity are techniques that employ attribute-based elicitation, asking people to specify weights for how much they care about a certain attribute (e.g., price). Because they can be challenging to use for non-domain experts, these attribute-based elicitations are mostly used in expert systems [19]. At the end of this complexity spectrum are conversational systems that can to elicit preferences as natural language expressions [8, 22], but the question of how to leverage these preferences for ML is an open problem [4, 16]. In our work, we limit our comparison to category-based, item-based and personality-based elicitation techniques because those are both well-established and well-suited for non-expert users.

Another dimension of preference elicitation techniques is whether they target a static or dynamic experience. In dynamic techniques based on active learning, people go through multiple rounds of providing ratings [32, 38]. Critiquing methods iteratively refine their current recommendations according to feedback they received. The work in [26] compares different dynamic experiences. The authors found having people self-specify items to rate increased loyalty with the system, even though it took longer than other onboarding experiences.

Finally, Pommeranz et al. [31] underline the importance of preference elicitation, articulation and representation in recommending systems. One design recommendation resulting from their set of studies is to onboard people with a static interest instrument because it was generally preferred over more dynamic experiences. For this reason, we focus on static experiences in this work, but discuss various extensions in Section 5.

3 NEED-FINDING STUDY

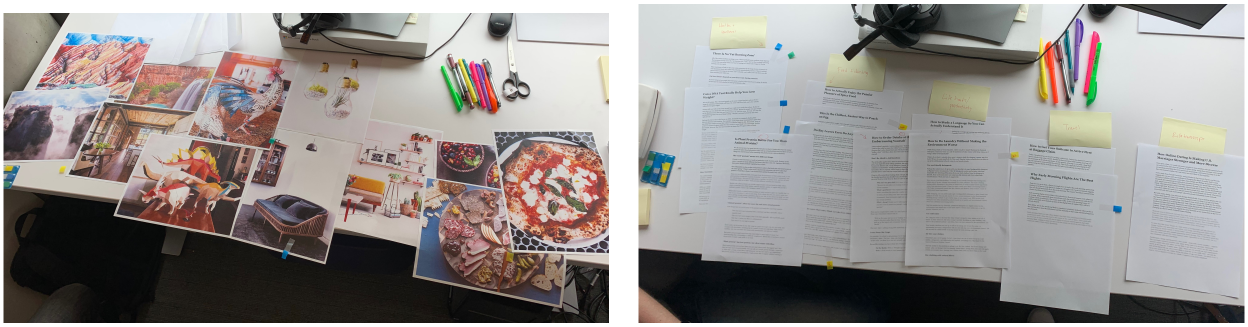
This section describes the need-finding study we carried out to better understand the rationale behind user preferences on common consumer content like text articles and images. Our goal was to understand how people justify and explain their choices in these settings and to distill a general set of patterns from these explanations. In Section 4, we will then illustrate how this set of patterns can be used to design a better onboarding experiments.

3.1 Participants

We recruited 23 (12 female, 11 male) participants through email lists at a large software technology company. The set of participants included individuals with diverse backgrounds such as research, design, and engineering. 20 of the Participants had a bachelor's or higher degree. Their median age bracket was 25-35 years.

3.2 Procedure and Tasks

After an initial demographics survey, we asked participants to complete two tasks, each requiring participants to select at least three



(a) Image selection task. Here, the participant arranged the images in overlapping thematic groups.

(b) Article selection task. The participant organized articles thematically in groups and added notes to represent each theme.

Figure 1: Snapshots of participants' working spaces during the need-finding study.

items from a set of 12 (either 12 online articles or 12 images) that they preferred the most while thinking-aloud. For each domain, we agreed on a set of four general interest categories and chose three items per category. For articles, we contextualized the user task as bookmarking articles for later reading. We presented articles as printed pages where only the first three paragraphs were legible to simulate the common activity of glancing at article snippets and then decided to return to read more later. For images, the task was contextualized as "liking" or pinning an image on social networks such as Instagram or Pinterest. We again presented images in physical form as printed pages in color. We balanced the presentation order of tasks across participants.

The main task for both domains was for participants to select their preferred items. Our intention was to give participants latitude regarding how they wanted to go through the items and express their decisions. To this end, we allowed participants to review, sort, group, or markup items in any way they liked. We also provided them with a set of common office supplies such as markers, highlighters, tape, post-it notes, and tags which they could use freely, such as for taking notes, tagging or grouping items, or marking up the items themselves.

After participants made their selections, we then asked them to narrow their set down to three items if they exceeded that number – we did this to make the choice task non-trivial, and to time-bound the follow up questions. For each item in the final set, we then asked participants "What are your main reasons for keeping this item?". To obtain more detailed responses, for each reason given we then asked:

- Are you just interested in <reason> this particular item or is this part of a more general interest?
- If we were to show you other articles/images about <reason>, would you also be interested?
- If not, what else would an article/image need for you to be interested in it?

After each task, we asked participants to fill out a qualitative survey about their experience and their perceived effort – based on the the Nasa TLX instrument. Overall, the study took approximately one hour and participants received a $25 cafeteria voucher for their time.

3.3 Items

Two of the authors independently collected a set of 12 images and 12 articles each from general interest websites and applications (e.g. Pinterest, unsplash.com, lifehacker.com). To have this dataset elicit a rich responses, chose the items with respect to the following criteria:

- *Diverse.* The set of items should be diverse, potentially impacting the variety of explanations we would obtain.
- *General.* Each item should be agnostic to people's backgrounds so that all items can be considered.
- *Attractive.* The items should be similarly attractive to minimize popularity effects.

We started by agreeing on a set of categories to choose items from (images: food, travel, design, crafts; articles: food, travel, lifestyle, health). We then independently gathered items, and iteratively discussed and eliminated items until we reached consensus. We did not aim for a comprehensive understanding of preferences but rather wanted to distill a set of rationales that would arise in everyday browsing of websites that target a general audience. Figure 1(a) and Figure 1(b) illustrate the materials as being used during the study.

3.4 Preference Explanations Codes and Themes

We examined our participants' responses for the items they selected (3 per task) to identify common preference explanation themes and articulation patterns using a grounded theory methodology [37]. All the authors of this paper (three in total) coded and discussed partially overlapping subsets of participant responses until the codes stabilized. During the coding, we sought to identify common patterns that could inform the design of a general interest preference elicitation instrument. We designed the codes to be mutually exclusive, and dimensions to be independent following best practice guidelines [6]. Through this process, we arrived at a set of preference explanation codes, describing two separate dimensions of a preference explanation. The first dimension describes how people connected to an item and found it to be relevant, while the second dimension captured the temporal dimension of relevance. Again,

	frequency			kappa
	overall	articles	images	overall
R1: actions, choices, biases	0.83	0.95	0.70	0.71
R2: item attributes	0.57	0.48	0.68	0.58
R3: people	0.12	0.13	0.11	0.80
R4: emotions, associations, curiosity, and imagination.	0.54	0.32	0.77	0.55
T1: started & ended in the past	0.12	0.06	0.18	0.60
T2: started the in past, ongoing	0.79	0.94	0.64	0.56
T3: in-the-moment	0.73	0.56	0.91	0.52

Table 1: Codes frequency and interrater agreement scores as Fleiss' kappa.

codes in each dimension are mutually exclusive, so that each explanation will have exactly two codes, one for its temporal and one for its relevance dimensions.

R1 - (Relevance) Actions, choices, biases. The item at hand is relevant because it reflects concrete, deliberate actions, choices, or biases that a person has. E.g., past or current hobbies, or plans for the future. This means that the relationship focuses on the person itself.

R2 - (Relevance) Item attributes. The item is relevant because of specific aspects of the item such as its content, structure, or interpretation. E.g., explanations that focus on the item, such as an object that catches the eye, or the tone of an article.

R3 - (Relevance) People. The item is relevant because it connects to people one knows. E.g., items that one would like to share with others, or that are about things one would like to do with others.

R4 - (Relevance) Emotions, associations, curiosity, and imagination. The item is relevant because of a felt reaction that it causes, which is typically hard to articulate and goes beyond the item. E.g., vague or experiential descriptions such as "this is beautiful", "I can picture myself there", or "reminds me of home".

For the temporal aspect, we grouped explanations according to when the underlying relationship would start and end.

T1 - (Temporal) Started & ended in the past. The explanation involved an event, interest or activity that has passed. E.g., recalling a past job.

T2 - (Temporal) Started in the past, ongoing. The explanation involved ongoing activities or behaviors. E.g., a current interest in a topic such as cooking, or an activity such as running.

T3 - (Temporal) In-the-moment or not started yet. This code captures immediate reactions as well as plans for the future. E.g., judgments or opinions such as "this looks rustic", "I'd like to go there", or questions such as "How was this made?".

Using these codes, we can now code each reason given along two dimensions. For example, P23 explained their selection like so: "*And this one is very beautiful <R4; T3>. I love waterfalls <R1; T2>, so that's really pretty <R4; T3>. If I was making a Pinterest board about different scenery or different places I want to visit, I would save that one because it very beautiful <R4; T3>.*". P15's reasons for selecting a photo depicting origami is another example of how we can use our themes to capture a person's explanations: "*But these, then I was like, were they glued together? <R4; T3> I mean there's all these pieces, is there some sort of three dimensional jigsaw puzzle? <R4; T3> If so how does it attach? <R4; T3> I like jigsaw puzzles and sometimes you can do these three dimensional ones which are kind of challenging <R1; T2> . And I mean who doesn't like dinosaurs, they're really interesting. <R1; T2>*" (P15). We expect that some correlation will exist between codes across these two themes, however our grounded analysis found this main subdivision expressive and leading to clear coding.

3.5 Findings

Table 1 illustrates the code's frequency by item type as well as the associated interrater agreement scores, which range from moderate to substantial. It is interesting to see the differences in frequency between images and text across some codes as they can reveal intrinsic characteristics of the medium and hint at what type of concepts people want to express for a particular type of document. For example, R4 (*emotions…*) occurred at least twice as often with images than articles. This seems reasonable, as imagery can be processed much quicker than text . These difference in frequencies underscores an opportunity to give users a rich language along these codes to express image preferences. Conversely, R1 (*actions*) and T2 (*started in the past, ongoing*) were at least 25% more frequent with articles than images. This makes sense as articles often articulate specific concepts people connect to and implies that the selection of articles is more informed by deliberate processes. These frequency differences suggest that a language to express preferences in articles should support those codes in greater detail.

While participants used different strategies to select preferred items, most engaged in a two- or three-step process where the first one or two passes involved browsing and then grouping the items, while the final step involved selection. These strategies are known as *consideration set formation* strategies in the context of making a purchase decision in marketing [12]. Initial groupings were often around preferences themselves (e.g, "yes", "no", and "maybe") or were topical (e.g., "food", "sports", "furniture", "health", etc.). When groupings were topical, selections always spanned multiple topics suggesting a desire for diversity within relevant recommendations. Few participants marked up the items themselves, choosing instead to spatially arrange or pile items as they made their selections.

3.5.1 Questionnaire Results. Regarding the questionnaire we gave people, on a Likert scale of 1-5, participants thought that it was important that a website/app catered its content to their preferences ($M = 3.95$, $SE = 0.14$). Similarly, participants expressed that they would be willing to put in effort if it helps the website/app learn their preferences quicker ($M = 3.31$, $SE = 0.16$).

3.6 Limitations and Opportunities

The purpose of this study was to distill how people articulate their choice of a preferred item. In particular, what are the higher-level building blocks used in their explanations? For our study, we made deliberate choices that scope its outcomes to the case of images and text in the context or collecting tasks. In the next sections, we use the result of our study to inform the design of a preference elicitation instrument for scenarios with a similar context.

A limitation of our need-finding study is we used a specific set of articles and images. While this set generated rich responses from people and lead to a set of codes and themes, it is unclear to what extent these results shaped can generalize to other domains. However, we propose that the same grounded methods we used in this study can and should be applied in other scenarios. We further strongly believe that the high-level *relevance* and *temporal* dimensions are simple concepts to help the characterization of preferences beyond the two domains we studied.

From this study, we identified the following concrete opportunities for improving preference elicitation:

- **Leveraging decision strategies.** We observed that many participants follwed a two-stage process where they first grouped items semantically into categories, and then select their favorites with more attention, similar consideration set formation [3]. However, current interfaces do not provide explicit support for such strategies. There is an opportunity to not only support people but also elicit richer feedback during this process, for example via shortlists where the UI supports compiling a temporary list of potentially relevant items [36].
- **Understanding blind spots.** Our findings revealed that people expressed their preferences differently for images than for articles. People responded emotionally (hard to articulate concepts) to images far more often. This suggests that for images, user-generated descriptions such as tags may be necessary to capture emotional aspects that go beyond current content understanding.
- **Designing better onboarding experiences.** Many current systems provide insufficient support during onboarding for people to express their preferences along the dimensions that we uncovered in our study, for example along R1. As we show in the case study in the next section, adding more support for these preference dimensions via a tailored onboarding questionnaire results in accurate recommendations under low effort.

4 CASE STUDY: DESIGNING AN ELICITATION INSTRUMENT FOR AN ONBOARDING QUESTIONNAIRE

We now illustrate how to use the taxonomy from our need-finding study to inform the design of a preferences elicitation instrument consisting of a set of descriptive dimensions people can use to express themselves. To this end, we map the subset of codes that we expect to generalize well to a set of questions. We want this set to be expressive, yet small so that the instrument is not time-consuming.

Started & ended in the past was the least frequent code (12%), thus we ignore it. The remaining *temporal* codes can be combined with *relevance* ones to define distinct and coherent preference dimensions. We map *actions + started in the past, ongoing* into an *ongoing* dimension. This combination captures activities and things that people like doing. We map *choices + in-the-moment* into an *aspirations* dimension. This combination captures concrete goals that people have. We map *biases + in-the-moment* into an *inspirations* dimension. This combination captures things that inspire people and have the potential to be concrete. Finally, we map *item attributes* directly to a *descriptive* dimension. This is a dimension that can generalize well across different documents. This left us with the following four dimensions that form the basis of our preference elicitation instrument: *descriptive*, *ongoing*, *inspirations*, and *aspirations*. We left out R3 (*people*) and R4 (*emotions*) in the design of our instrument because we deemed them too specific to generalize from, but plan to investigate them in future work. For each of the four dimensions, we then looked at the list of responses from participants, and selected up to 10 response options that be able to summarize participants' overall answers best. Response options had to be mentioned at least twice to be considered, and we merged synonymous responses. Table 2 describes our final instrument which we call *Tell Why* or *TellY* for short. To implement TellY, we rendered responses as checkboxes with an additional *None of the above* option as shown in Figure 2.

4.1 Crowdsourced Validation Study

In this section, we describe a between-subjects user study we conducted to compare TellY to conventional preference elicitation instruments with respect to user effort and recommendation accuracy in the context of the types of tasks and documents we observed in our needs-finding study. The domains were the same as the ones in the need-finding study. We examine the following hypotheses:

H1. TellY requires comparable or less user effort compared to conventional instruments.

H2. TellY enables more accurate machine-learned recommendations compared to conventional instruments.

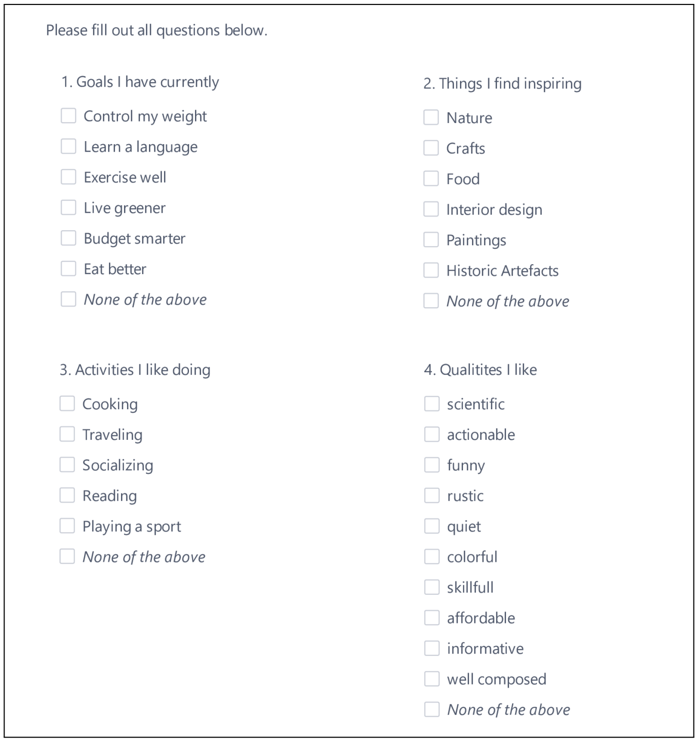
We measure the different elicitation instruments on three key constructs: effort, privacy, and perceived informativeness.

4.2 Conditions

We compare TellY to three common preference elicitation instruments from the literature and current systems. This led to the following four instrument conditions:

dimension	prompt	options		corresponding codes
descriptive	Qualities I like	• scientific • actionable • funny • skillful • affordable	• rustic • quiet • colorful • informative • well composed	R2 (+ T1-3)
ongoing	Activities I like doing	• cooking • traveling • playing a sport	• socializing • reading	R1+T2
inspirations	Things I find inspiring	• nature • crafts • food	• interior design • paintings • historic artefacts	R1+T3
aspirations	Goals I have currently	• control my weight • learn a language • exercise well	• live greener • budget smarter • eat better	R1+T3

Table 2: The TellY preference elicitation instrument. For each dimension, we provide a vocabulary for users to express how a dimension relates to their preferences.



Please fill out all questions below.

1. Goals I have currently

□ Control my weight
□ Learn a language
□ Exercise well
□ Live greener
□ Budget smarter
□ Eat better
□ *None of the above*

2. Things I find inspiring

□ Nature
□ Crafts
□ Food
□ Interior design
□ Paintings
□ Historic Artefacts
□ *None of the above*

3. Activities I like doing

□ Cooking
□ Traveling
□ Socializing
□ Reading
□ Playing a sport
□ *None of the above*

4. Qualitites I like

□ scientific
□ actionable
□ funny
□ rustic
□ quiet
□ colorful
□ skillfull
□ affordable
□ informative
□ well composed
□ *None of the above*

Figure 2: The TellY instrument as shown to participants.

Category-based. Many online systems allow people to select categories they are interested to help personalization. We compiled our list of categories by reviewing current news apps, such as Hummingbird and Apple News. The final list included the following twelve categories: Politics, Science and Tech, Entertainment, Sports, DIY and Hobbies, Art and Design, Lifestyle, Business, Health and Fitness, Fashion, Travel and Outdoors, Food and Dining.

Item-based. Implementing an item-based onboarding experience, people were asked to rate a set of twelve items (six articles, six images) with respect to whether they perceived an item to be generally interesting or not.

- **Personality-based.** Another instrument that has been used to inform personalization is personality quizzes [13, 14]. We use the popular Ten Item Personality Measure (TIPI) [10] of the Five-Factor model.
- **TellY.** This is the instrument that we derived from the formative study responses shown in Table 2. It asked four questions with 5-10 answer options each.

4.3 Participants

We recruited 518 (63% male, 37% female, and <1% chose not to identify) participants from Amazon Mechanical Turk. Participants had to have an approval rate of at least 95%, a modern browser with JavaScript, and be from a US-based location. The mean age was 36.2 years ($SD = 10.4$).

4.4 Procedure and Task

Participants were randomly assigned to one of the four instrument conditions (category, item, personality, ours). First, participants had to complete the instrument corresponding to their condition. After that, we asked them to complete a survey about the effort, privacy and perceived informativeness of the instrument with an attention check question.. We then prompted participants to complete two consecutive curation tasks with 12 items each, one task comprising only images and the other one comprising only articles. Analogous to the setup of the need-finding study we asked them to spend at least one minute on each task, bookmarking the three or more most interesting items that they might like to return to later. We used the same set of articles as in the need-finding study, and a similar set of images as in the need-finding study that had no usage or license restrictions. We randomized the order of the curation tasks (articles or images), as well as the position of each item on the page to guard against ordering effects. Participants were only allowed to proceed if they met the task requirements (more than one minute spent and three or more items bookmarked). Finally, they filled out an exit survey that asked for their demographic information as well as asked them to label two previously interacted items (one was bookmarked, the other was not) as an attention check. Participants were paid $2.00 for the successful completion of the experiment which had an average completion time of 5.5 minutes.

4.5 Analysis

From the 518 completed experiment sessions, we excluded all sessions from the analysis in this and the next section where participants failed to correctly answer the attention check (8 sessions). The number of participants for each condition ranged between 122 and 135. We used a subset of the Nasa's TLX measures (mental demand, effort, frustration level), and averaged them to create a single score for effort. For privacy concerns, we asked: "I would feel comfortable sharing my background survey with another person.". Lastly, we also inquired about how informative people felt the information was by asking: "Based on my background survey, a person or machine should be able to tell which articles or images I would like." All responses used a 7-point Likert scale, with 1 corresponding to "strongly disagree". For all continuous-valued responses, we ran one-way ANOVA analyses. Similarly, we tested for differences in ordinal variables via a Kruskal-Wallis test. If we were able to reject the null hypothesis, we followed up with a Tukey test for testing for differences between pairs, assuming a Normal distribution.

4.6 Results

method	effort	privacy	informativeness
personality	2.30 ± .19	5.18 ± .30	3.29 ± .26
category	2.27 ± .20	5.78 ± .22	**4.64 ± .26**
item	**2.53 ± .19**	**5.97 ± .22**	4.51 ± .23
TellY	2.26 ± .18	5.32 ± .26	3.92 ± .26

Table 3: Qualitative user-centric measures aggregated from survey responses. All measures were derived from 7-point Likert scales (range 1-7). The highest number of each column is in bold.

Table 3 shows the results for the three key constructs we tested: effort, privacy, and perceived informativeness. Starting with effort, participants reported low effort with averages being well below 3. People reported the highest effort under the item-based instrument, and the ANOVA showed significant differences between the four conditions ($p < 0.05$). The post-hoc test revealed significant differences between the category-based instrument and the item-based one ($p < 0.05$). The lowest effort was reported under TellY, although it is not significantly different from the other three conditions. Although not shown in the table, this correlates roughly with the median completion times for all instruments (personality = 29s, category= 16s, item = 55s, TellY = 38s). In short, the item-based instruments mark the upper end of the observed effort spectrum.

When assessing privacy concerns, we use higher scores to indicate a lower degree of concern. Generally, participants felt only mildly concerned about the information that they shared in the instrument, with the item-based instrument showing the least degree of concerns ($p < 0.01$). The category-based, personality-based instruments and TellY had slightly higher reported levels of concern, but were statistically indistinguishable from each other.

Finally, considering perceived informativeness, the ANOVA with a subsequent post-hoc test revealed that perceived informativeness was greatest for the item-based and category-based instruments, followed by TellY and eventually the personality-based instrument ($p < 0.001$ for all tests). Perhaps the most surprising is that people felt that a category-based instrument would be able to predict their interests well. However, one explanation is that people are very familiar with specifying preferences in categories which may bias their assessment.

4.7 Impact on ML performance

We now compare the four conditions in our large-scale study with respect to how accurately the information collected by each instrument can generate recommendations.

To assess ML performance, we set up the following supervised multilabeling task. Given the information provided in each instrument, predict the set of articles and images that the same person would bookmark during the curation tasks later in the session. With

12 images and 12 articles to choose from, this corresponds to 24 binary labeling decisions. We assess and report performance via the macro-averaged F1 score [34] of the predictions on the held-out test set. The F1 score is particularly useful in scenarios where label distributions are imbalanced which is the case in our dataset where only 35%-38% of all items were bookmarked (i.e., had positive labels). Ordinal responses from Likert scales are encoded as integer features and binary responses from checkboxes are encoded as binary features. We split the participant data into training (80%) and test sets (20%). To enable a fair comparison between the different conditions, we make sure that the participant data of each condition has an equal number of samples by selecting $N = 122$ random participants. To guard against inferior performance due to choosing a suboptimal algorithm, we exhaustively search through a large array of widely established classification algorithms (k-nearest neighbors, logistic regression, naive Bayes, Support Vector machines with linear or radial kernels). We perform five-fold cross-validation on the training set to choose the best performing algorithm and its parameters. We retrain with the best configuration on the entire training set and report performance on the test set. Finally, due to sampling variance from the small dataset size, we repeat this process 25 times and report the mean performances with their double standard errors. For reference, we also included a popularity-based baseline which always predicts the three articles and images that were bookmarked most often across all participants.

method	all	articles	images
popularity baseline	0.08 ± .00	0.09 ± .00	0.08 ± .00
personality	0.26 ± .01	0.22 ± .01	0.29 ± .01
category	0.30 ± .01	0.29 ± .01	0.30 ± .01
item	0.30 ± .01	0.27 ± .01	0.34 ± .01
TellY	**0.34 ± .01**	**0.31 ± .01**	**0.36 ± .01**
TellY w/o background	0.29 ± .01	0.29 ± .01	0.30 ± .01
TellY w/o descriptive	0.28 ± .01	0.26 ± .01	0.30 ± .01
TellY w/o inspirations	0.28 ± .01	0.24 ± .01	0.32 ± .01
TellY w/o aspirations	0.28 ± .01	0.28 ± .01	0.28 ± .01

Table 4: Predictive performance of different instruments measured by the macro-averaged F1 score on a separate set of test users. Statistics were computed over a set of 25 repetitions.

Table 4 presents the results of the analysis. Starting with the popularity baseline, we can see how its predictive performance is quite low with an overall F1 score of 0.08. This is in contrast to most typical recommendation scenarios where its performance is quite competitive. This indicates that there are large individual differences in what items people chose in the curation tasks. This, in turn, makes the prediction task harder since one cannot rely on overall item popularity. The second interesting observation is that performance on articles is generally lower than on images, suggesting that the former is the more challenging prediction domain. Comparing the four instrument conditions, we can see that the personality-based instrument resulted in the lowest scores among the four. This is in line with our intuitions because it lacks the ability to capture critical topical information. The category-based and item-based elicitation methods yield comparable overall performance, with the category-based method being stronger on articles, and the item-based method winning in the image domain. Finally, TellY outperforms all other instruments, independent of whether we consider performance on all items, or split by item type (articles or images).

We also conducted an ablation study with our elicitation method to examine which parts of TellY were most important. The results are shown in the bottom part of Table 4. Overall, eliminating any of the questions resulted in a drop in performance, with most questions showing similar scores between 0.28 and 0.29. This implies that the questions are all viable to good predictive performance, and none of them was fully captured by the remaining three questions.

4.8 Summary

Regarding the research hypothesis *H1*, we did see that TellY required a similar effort from people as conventional methods. Given that TellY asked more questions than category-based elicitation, finding that it poses similar effort to people implies that not the sheer quantity of questions of an instrument is important, but also their quality. This is encouraging because it allows us to more make more flexible trade-offs during the instrument design process.

We did not find that people thought that the information they provided through TellY would be significantly better than the information in item-based or category-based at predicting which images or articles they might like. We believe that this is partially due to people's familiarity with these common onboarding experiences. We also note that perceived informativeness does not imply or is equivalent to ML performance as our results show.

Overall, we saw that TellY was able to provide improved machine-learned recommendations when compared to conventional elicitation methods. This allows us to confirm the research hypothesis *H2*. Moreover, we found it improved recommendation performance on both images and articles, implying that it does indeed capture certain domain-independent factors.

4.9 Discussion

Studying this particular instrument in an online setting with more than 500 participants revealed the following insights:

- Item-based elicitation is perceived to ensure the most privacy but at a cost of more user effort and lower recommendation performance.
- Category-based elicitation required the least effort and provides comparable perceived privacy and recommendation performance as an item-based recommendation. Hence, it dominates item-based elicitation in this setup.
- Personality-based had the lowest recommendation performance and perceived privacy in our study, we would caution against this method in practice.

Figure 3 shows the trade-offs that the different techniques are subject to on the user effort vs. recommendation performance spectrum. For reference, we also included the popularity-based recommender, marking the low end of required effort but at the cost

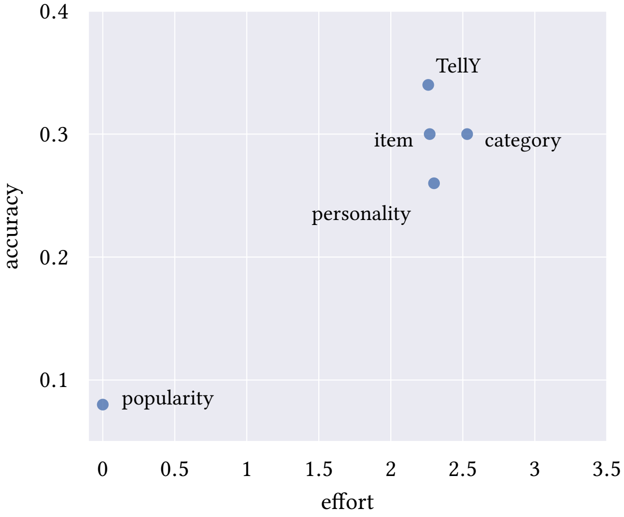


Figure 3: Different elicitation methods fall onto different parts of the accuracy vs. effort spectrum. TellY provides highest accuracy under low effort.

of substantially lower accuracy. Related to recommendation performance is the finding is that even though people *perceived* the category-based or item-based instruments to be most informative for recommendations, our machine learning results showed that these intuitions were incorrect. This disconnect between what information people believe machine learning systems can effectively learn from and what they are actually capable of learning from is an interesting avenue for future investigation.

While our case study focuses on the usage of TellY during an onboarding or cold-start scenario, we hypothesize that it may also be useful beyond this scenario as a way to capture the changing nature of a person's preferences over time. For example, future work may examine the use of TellY's prompts and options to explain why a particular recommendation is being presented (e.g., "Recommended because you care currently trying to live greener"), potentially providing a seamless opportunity for people to revisit and adjust their preferences over time. This extra layer of control may also impact user trust in their recommender system.

During our case study, we instantiated TellY in the form of a static questionnaire. Future work may explore experiences that are more dynamic and, in turn, more efficient. For example, a preference elicitation instrument could present users with possible personas to choose from to start (e.g., "the weekend warrior", or the "politics junkie"), derived from common clusters of user responses, and then edit accordingly. While our ablation analysis suggests that each of TellY's prompts add to its performance capabilities, another potential method for reducing costs may be to automatically rank order or present only a subset of the most popular prompts and options to start, providing access to others via progressive or hierarchical disclosure.

Our need-finding study also revealed many articulation patterns that we did not include in TellY because we either found them difficult to elicit via a multi-choice prompt or because we did not believe the patterns could generalize to other items via machine learning. For example, it is hard to extrapolate from a sharing pattern such as "My friend should read this" without additional information about what aspects make an item shareworthy (which we believe are captured better with the prompts and options we included with TellY) and with whom. Future research may examine other uses of such articulation patterns to improve recommender systems, perhaps augmented with social network information. Moreover, we would like to explore how to infer salient patterns or options automatically from tagged content, such as images.

5 CONCLUSIONS

Our paper contributes to the space of techniques for eliciting user preferences – aiming at increasing the level of control people have over their recommended content. Through a need-finding study, we study how people articulate preferences during curation tasks with text articles and images. Our observations reveal the different dimensions people rely on when expressing preferences about articles and images. Among the different opportunities this study opens, we chose to explore how to use our need-finding study results to build a rich onboarding preference elicitation instrument, TellY. In our case study with TellY, we compare it with four commonly used preference onboarding methods. Overall, we find that TellY leads to more accurate predictions than other methods, while maintaining user satisfaction under low effort.

These results encourage us to push towards research on giving people more agency over the information they want the recommender system to use. There is power in large numbers of implicit, low-bandwidth signal streams – such as views and clicks. We believe, however, that one can complement this information by people's active participation and the rich signals they provide. This framing of human-AI collaboration can lead to increased accuracy, as well as improved qualitative metrics in recommender systems.

REFERENCES

[1] Xavier Amatriain, Josep M Pujol, Nava Tintarev, and Nuria Oliver. 2009. Rate it again: increasing recommendation accuracy by user re-rating. In *Proceedings of the third ACM conference on Recommender systems*. ACM, 173–180.

[2] Naveen Farag Awad and Mayuram S Krishnan. 2006. The personalization privacy paradox: an empirical evaluation of information transparency and the willingness to be profiled online for personalization. *MIS quarterly* (2006), 13–28.

[3] Lee Roy Beach. 1993. Broadening the definition of decision making: The role of prechoice screening of options. *Psychological Science* 4, 4 (1993), 215–220.

[4] Toine Bogers and Marijn Koolen. 2017. Defining and supporting narrative-driven recommendation. In *RecSys*. 238–242.

[5] Carolyn Brodie, Clare-Marie Karat, and John Karat. 2004. Creating an E-commerce environment where consumers are willing to share personal information. In *Designing personalized user experiences in eCommerce*. Springer, 185–206.

[6] John L Campbell, Charles Quincy, Jordan Osserman, and Ove K Pedersen. 2013. Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research* 42, 3 (2013), 294–320.

[7] Ben Carterette, Paul N Bennett, David Maxwell Chickering, and Susan T Dumais. 2008. Here or there. In *European Conference on Information Retrieval*. Springer, 16–27.

[8] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 815–824.

[9] Bruce Ferwerda, Emily Yang, Markus Schedl, and Marko Tkalcic. 2015. Personality Traits Predict Music Taxonomy Preferences. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '15)*. ACM, New York, NY, USA, 2241–2246. https://doi.org/10.1145/2702613.2732754

[10] Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in personality* 37, 6 (2003), 504–528.

[11] Tom Hope and Dafna Shahaf. 2018. Ballpark crowdsourcing: The wisdom of rough group comparisons. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 234–242.

[12] J.A. Howard. 1977. *Consumer behavior: application of theory*. McGraw-Hill.

[13] Rong Hu and Pearl Pu. 2009. A Comparative User Study on Rating vs. Personality Quiz Based Preference Elicitation Methods. In *Proceedings of the 14th International Conference on Intelligent User Interfaces (IUI '09)*. ACM, New York, NY, USA, 367–372. https://doi.org/10.1145/1502650.1502702

[14] Rong Hu and Pearl Pu. 2011. Enhancing collaborative filtering systems with personality information. In *Proceedings of the fifth ACM conference on Recommender systems*. ACM, 197–204.

[15] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *ICDM*. 263–272.

[16] Jie Kang, Kyle Condiff, Shuo Chang, Joseph A Konstan, Loren Terveen, and F Maxwell Harper. 2017. Understanding how people use natural language to ask for recommendations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 229–237.

[17] Diane Kelly and Jaime Teevan. 2003. Implicit feedback for inferring user preference: a bibliography. In *SIGIR Forum*, Vol. 37. 18–28.

[18] Daniel Kluver, Tien T Nguyen, Michael Ekstrand, Shilad Sen, and John Riedl. 2012. How many bits per rating?. In *Proceedings of the sixth ACM conference on Recommender systems*. ACM, 99–106.

[19] Bart P Knijnenburg, Niels JM Reijmer, and Martijn C Willemsen. 2011. Each to his own: how different users call for different interaction methods in recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. ACM, 141–148.

[20] Bart P Knijnenburg, Martijn C Willemsen, and Stefan Hirtbach. 2010. Receiving recommendations and providing feedback: The user-experience of a recommender system. In *International Conference on Electronic Commerce and Web Technologies*. Springer, 207–216.

[21] Shyong Lam, Dan Frankowski, and John Riedl. 2006. Do you trust your recommendations? An exploration of security and privacy issues in recommender systems. *Emerging trends in information and communication security* (2006), 14–29.

[22] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. In *Advances in Neural Information Processing Systems*. 9725–9735.

[23] Ting Li and Till Unger. 2012. Willing to pay for quality personalization? Trade-off between quality and privacy. *European Journal of Information Systems* 21, 6 (2012), 621–642.

[24] Yabing Liu, Krishna P Gummadi, Balachander Krishnamurthy, and Alan Mislove. 2011. Analyzing facebook privacy settings: user expectations vs. reality. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM, 61–70.

[25] Robert R. McCrae and Oliver P. John. 1992. An Introduction to the Five-Factor Model and Its Applications. *Journal of Personality* 60, 2 (1992), 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-6494.1992.tb00970.x

[26] Sean M McNee, Shyong K Lam, Joseph A Konstan, and John Riedl. 2003. Interfaces for eliciting new user preferences in recommender systems. In *International Conference on User Modeling*. Springer, 178–187.

[27] Hien Nguyen and Peter Haddawy. 1999. The decision-theoretic interactive video advisor. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 494–501.

[28] Tien T Nguyen, Daniel Kluver, Ting-Yu Wang, Pik-Mai Hui, Michael D Ekstrand, Martijn C Willemsen, and John Riedl. 2013. Rating support interfaces to improve user experience and recommender accuracy. In *Proceedings of the 7th ACM conference on Recommender systems*. ACM, 149–156.

[29] Syavash Nobarany, Louise Oram, Vasanth Kumar Rajendran, Chi-Hsiang Chen, Joanna McGrenere, and Tamara Munzner. 2012. The design space of opinion measurement interfaces: exploring recall support for rating and ranking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2035–2044.

[30] David M Pennock, Eric Horvitz, Steve Lawrence, and C Lee Giles. 2000. Collaborative filtering by personality diagnosis: A hybrid memory-and model-based approach. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*. 473–480.

[31] Alina Pommeranz, Joost Broekens, Pascal Wiggers, Willem-Paul Brinkman, and Catholijn M Jonker. 2012. Designing interfaces for explicit preference elicitation: a user-centered investigation of preference representation and elicitation process. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 357–397.

[32] Al Mamunur Rashid, Istvan Albert, Dan Cosley, Shyong K Lam, Sean M McNee, Joseph A Konstan, and John Riedl. 2002. Getting to know you: learning new user preferences in recommender systems. In *Proceedings of the 7th international conference on Intelligent user interfaces*. ACM, 127–134.

[33] Paul Resnick and Hal R Varian. 1997. Recommender systems. *Commun. ACM* 40, 3 (1997), 56–59.

[34] C. J. Van Rijsbergen. 1979. *Information Retrieval* (2nd ed.). Butterworth-Heinemann.

[35] Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. 2002. Methods and metrics for cold-start recommendations. In *SIGIR*. 253–260.

[36] Tobias Schnabel, Paul N Bennett, Susan T Dumais, and Thorsten Joachims. 2016. Using shortlists to support decision making and improve recommender system performance. In *Proceedings of the 25th International Conference on World Wide Web*. 987–997.

[37] A.L. Strauss and J.M. Corbin. 1990. *Basics of qualitative research: grounded theory procedures and techniques*. Sage Publications. https://books.google.com/books?id=nvwOAQAAMAAJ

[38] Mingxuan Sun, Fuxin Li, Joonseok Lee, Ke Zhou, Guy Lebanon, and Hongyuan Zha. 2013. Learning multiple-question decision trees for cold-start recommendation. In *Proceedings of the sixth ACM international conference on Web search and data mining*. 445–454.

[39] Kristen Vaccaro, Dylan Huang, Motahhare Eslami, Christian Sandvig, Kevin Hamilton, and Karrie Karahalios. 2018. The Illusion of Control: Placebo Effects of Control Settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 16, 13 pages. https://doi.org/10.1145/3173574.3173590