How Does the Number of Participants Influence the Generalizability of Interactive IR User Studies?

Jiepu Jiang
jiepu.jiang@wisc.edu
University of Wisconsin-Madison
Madison, Wisconsin, USA

ABSTRACT

User studies are a widely used research method for Interactive information retrieval (IIR). However, the limited number of participants raises many concerns about the generalizability of its findings. The IIR community does not have a standard guideline regarding how many participants to recruit. We study how the number of participants influences the generalizability of research findings in IIR user studies using a simulation-based method. We reproduced Kelly et al. (2015)'s experiment with 500 participants from the Amazon Mechanical Turk. We then sampled experiment participants out of the 500 participants to simulate experimental results with fewer participants. We example whether the (simulated) experiments with fewer participants can come to the same findings as that with 500 participants. By comparing the results' difference between the smaller samples of participants and the original one, we conclude the suggested sample size to reduce the chances of generalizability at a certain level. Results suggest that different dependent variables require substantially different numbers of participants to ensure the findings are generalizable. Also, we found that the theoretical analysis and an experiment's type I and II errors may be underestimated. Our study provides guidelines for IIR research determining the number of participants.

CCS CONCEPTS

• **Computer systems organization → Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

datasets, neural networks, gaze detection, text tagging

ACM Reference Format:
Jiepu Jiang. 2018. How Does the Number of Participants Influence the Generalizability of Interactive IR User Studies?. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

IIR user studies often rely on inferential statistical methods to conclude if an independent variable has significant effects on a dependent variable. Popular independent variables in IIR studies include task attributes, search devices or interfaces, participants' traits, search environment, etc. The dependent variables usually include different search behaviors, search performance, and user experience measures. Depending on the problem and experimental design, we may use various inferential statistical methods such as t-test, ANOVA, and regression to examine effects. A common practice is to claim a significant effect if the adopted test gets a p-value below a selected threshold (alpha level). Although the community has criticized the overuse and misuse of p-value, significance tests and p-value are still the primary tools for concluding variables' effects.

Lab-based user studies often raise concerns about generalizability and reproducibility. For example, the lab environment and the search tasks may more or less deviate from the organic search scenarios. Also, lab-based studies are challenging to scale up and often only include a limited number of participants. Here we focus on the number of participants, a salient factor influencing the validity of lab-based user studies, in the specific context of Interactive Information Retrieval (IIR) research. Our purpose is to examine how the number of participants for an IIR user study influences the generalizability and reproducibility of the study's findings.

Many inferential statistics methods can examine the generalizability of a significance test's results. For example, in theory, the alpha level of a significance test determines the chances of type I errors (overclaiming an effect that does not exist). Also, many significance tests have power analysis to compute the number of participants needed to maintain the power (and equivalently, the chances of type II errors) at a certain level. However, such analysis may more or less diverge from the actual risks of errors because the observations rarely perfectly conform to the data assumptions of the statistical methods (e.g., normal distributions with equal variance). Also, such general analysis may not provide many insights into specific problems pertinent to IIR research.

Thus, we examine the generalizability of IIR user studies using a different approach. For an existing experiment, we randomly sample a smaller group of participants' observations from the collected data and examine the chances that the findings (such as performing a significance test to determine an effect) are consistent with the whole group of participants. We call such chances of inconsistencies generalization error rate in this paper. We use generalization error rate to complement the theoretical analysis of type I and II errors to examine IIR user studies. We reproduced an IIR user study by Kelly et al. [22]. We examined the influence of the number of participants on the generalizability of findings (whether IIR search

Woodstock '18, June 03–05, 2018, Woodstock, NY
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00
https://doi.org/10.1145/1122445.1122456

task complexity level has a significant effect on search behavior and user experience measures).

Our study has made the following contributions:

First, we have examined how generalization error rates based on an actual IIR experiment's data diverge from the estimated chances of type I and II errors in theory. We show that the chances of type I and II errors based on theoretical analysis (the adopted alpha level or power analysis) are underestimated. The extent of underestimation differs by variables in IIR studies.

Second, we suggest the number of participants needed to maintain the generalization error rates at a certain threshold based on the collected data. These numbers are helpful for future experimental designs using similar settings (e.g., using tasks of varied cognitive complexity levels developed by Kelly et al. [22]).

2 RELATED WORK

2.1 IIR User Studies

User studies are important methods for interactive information retrieval (IIR) research. But user studies often raise generalizability concerns due to the limited number of participants, especially in a lab-based setting. We reviewed 50 user studies from 48 full-length articles [1–4, 7–21, 23–27, 29–40, 42, 44–49, 51, 52, 54–56] published on SIGIR (2014–2018) and CHIIR (2017–2018), which provides a landscape to IIR user studies in recent studies. These studies included 47 lab-based user studies and three online ones.

The reviewed studies have included a very limited number of participants (mean 37.1, SD 15.9). The smallest number of participants is 10 [11, 51], and the largest sample size is 72 [40]. The majority of these studies recruited 20 to 55 participants, which casts doubts on the reliability and generalizability of these studies. Sakai [41] mentioned in a systematic review of SIGIR full papers and TOIS papers from 2006 to 2015 that: 1) a sample size of 28 participants is too small to investigate the user experience sub-scale in two systems; 2) the result from an experiment with 24 participants is underpowered in comparing two algorithms. Except for those research that focuses on evaluating individuals' behavior, a larger sample size is likely to make the results more representative of the entire population and get a more generalizable conclusion.

IIR studies' independent variables usually depend on the specific problem of interest. Among the reviewed 50 studies, search interfaces are one of the most widely employed independent variables, such as interactive multilingual search interfaces [29], interfaces with different snippet lengths [37], and intelligent assistants [25]. In addition to search interfaces, search task attributes are another frequently used independent variable in lab-based user studies, including different search task types [5, 28] and cognitive complexity levels [22], such as simple and complex [44], fact-finding and exploring [17, 23, 27, 35, 52, 54]. However, most studies have used different tasks and slightly different task dimensions, despite some existing work providing a standard task pool [22, 53].

IIR studies' dependent variable usually included various search behavior and user experience measures. Search behavior measures indicate the interactions between participants and the tested interfaces or systems, usually including the number of search queries, the number of clicks, and the number of information viewed by the participants. In addition, IIR studies also measure searchers' perceptions of the system or the search process, such as how engaged people feel while interacting with the system, to what extent they are satisfied with the searching experience and the solution, and whether they experience any difficulty during the interaction.

Reviewing the 50 IIR user studies in SIGIR and CHIIR articles helps us determine the reproducing experiment. Here we chose to reproduce Kelly et al.'s [22] experiment examining the effect of task cognitive complexity levels on search behavior and perceived task attributes before and post search. We apply our analytical approach to the reproducing experiment's data to suggest the number of participants needed for each dependent variable. This helps future researchers design experiments with similar settings and variables as many had followed Kelly et al.'s [22] tasks and used task cognitive complexity levels as an independent varible.

2.2 Stability of IR Experiments

The information retrieval community has studied the stability of Cranfield-style evaluation of retrieval experiments. For example, Buckley and Voorhees [6] introduced a methodology to evaluate the stability of IR experiments and suggested the optimal number of queries for web measurements. Voorhees and Buckley [50] also introduced a swap method to calculate the error rate of relevant documents between TREC data and gave the suggested topic set size for information retrieval experiments. Sanderson and Zobel [43] also introduced a method different from Voorhees and Buckley's methods [50] for a similar purpose. Sakai [41] introduced a Bootstrap-based method to investigate the sensitivity of Information Retrieval metrics for ranking a list of systems. Our method of examining the generalizability and reproducibility is conceptually similar to Voorhees and Buckley's error rates [50], except that we apply it to examine the consistencies of Interactive IR studies with limited sample size.

3 METHOD AND EXPERIMENT

3.1 Generalization Error Rate

Interactive Information Retrieval (IIR) user studies often rely on inferential statistical methods (e.g., t-test, ANOVA, and regression) to determine if an independent variable (IV) has significant effects on a dependent variable (DV). A common practice is to suggest a significant effect if the adopted significance test gets a p-value below a chosen threshold (alpha level).

We are interested in the effects of the number of participants on the generalizability of the findings using significance tests in IIR user studies. In theory, a significance test's alpha level determines the chances of type I errors (overclaiming an effect that does not exist), independent of the sample size (the number of participants). The chances of type II errors (missing an actual effect) decrease when the sample size increases. Many significance tests have mature methods to estimate the number of participants required to maintain the power of the tests (or, equivalently, to restrict the chances of type II errors) at a certain level. But do the theoretical estimates agree with the practice in IIR studies?

Here we take a different angle to look into this problem by examining an existing experiment's generalization errors. Suppose we have finished an experiment with N participants, and we trust

Table 1: What generalization error rate measures when the experiment agrees or disagrees with the truth.

		Truth	
		Sig. Effect	No Sig. Effect
Experiment	Sig. Effect	Type II	1 − Type I
	No Sig. Effect	1 − Type II	Type I

its outcome (whether an IV has a statistically significant effect on a DV based on a significance test). We may use this experiment's outcome as the ground truth to examine the chances of errors (inconsistencies) while conducting another experiment using the same setting with n participants. We define the chances of such errors (inconsistencies) as the error rate of generalizing the experiment to n participants, or **generalization error rate** for short.

When the outcome of the original experiment with N participants aligns with the truth, the generalization error rate estimates type I or II errors for experimenting with n participants. However, when the outcome of the original experiment is different from the truth, the generalization error rate estimates 1− type I or II errors, as Table 2 shows. But practically, we may expect such generalization error rates to provide an alternative estimate to type I and II error rates. If we can estimate it accurately, such generalization error rates can complement the theoretical estimate of type I and II errors in practice. Actual experimental observations more or less violate the data assumptions required for significance tests (e.g., are not perfectly normal distributions), which may make the true type I and II errors differ from the theoretical levels.

In practice, we can estimate a restricted but easy-to-compute form of the generalization error rate by sampling the size n sample from the N total participants (without replacement). Each size n sample is a subset of the total N participants. Suppose we sample n participants from N for many times (K) and record the number of times that the same significance test and alpha level come to the same results (k). We estimate the error rate as k/K.

This estimation of generalization error rate has some properties:

- It is easy to estimate, which makes it practically useful. It does not require multiple experiments to come to an estimate of the error rate.
- It underestimates the actual error rate of generalizing from sample size n to N. This is because the size n samples are subsets of the N total participants. Thus they are more similar to the original N participants than a random size n sample from the actual population, which underestimates the error rates.
- When n increases and approaches N, the size n samples are more similar to the N total participants and thus underestimate the error rates by a greater extent. Specifically, the estimated error rate is 0 when $n = N$.
- Assuming the significance test's outcome on the N total participants is correct, the restricted estimation of generalization error rate also provides a lowerbound for the type I and II errors.

Therefore, we expect this practical but biased (underestimated) estimation from a single experiment's observations to provide an easy-to-compute lowerbound to the actual generalization error rate and the experiment's type I and II errors. It is not perfect, but practically helpful (as it is easy to measure). Particularly, the estimation is more accurate when n is small.

Table 2: Kelly et al. [22] vs. our experiments.

	Kelly et al. [22]	**Our Experiment**
IV	task cognitive complexity (five levels)	
DV	11 search behavior measures, 11 pre-task questions, 13 post-task questions	
Environment	lab	crowdsourcing
Design	within-subjects	between-subjects
Sample Size	48 observations for each IV level	100 observations for each IV level

We use this restricted estimation of generalization error rates to examine an IIR user study's findings. In future sections, we use generalization error rates to refer to this restricted version.

3.2 Estimated Type I and II Error Rates

In addition to computing generalization error rates, we also use another approach to estimate type I and II errors (regarding a sample size n) based on an existing experiment's observations.

If we observe an IV has a significant effect on a DV in the experiment's N observations, we estimate a DV distribution for each IV level. Then, we randomly sample n data points of the DV for each IV level using that IV level's estimated DV distribution. We repeat the process many times (K) and record the number of times (k) that the IV does not show a significant effect on the sampled DV data points. We estimate the chances of type II error (missing an actual effect) regarding a sample size n as k/K.

Suppose an IV does not show a significant effect on a DV in the experiment's N observations. In that case, we estimate one single DV distribution using all the DV's observations from different IV levels. Then, we sample n data points of the DV for each IV level using the estimated DV distribution. We repeat the process many times (K) and record the number of times (k) that the IV shows a significant effect on the sampled DV data points. We estimate the chances of type I error (overclaiming an effect that does not exist) regarding a sample size n as k/K.

Note that for most parametric tests (e.g., t-test and ANOVA), if we estimate the DV distributions as Gaussian distributions, the estimated chances of type I error will approach the chosen alpha level. The estimated chances of type II error will approximate those derived from power analysis, i.e., power = 1 − type II error. However, we introduce this approach to reserve some flexibilities, e.g., we may model the DVs using other distributions (although we did estimate using Gaussian distributions in this paper), and we may use significance tests that cannot derive type II error rates analytically.

3.3 Reproducing Kelly et al.'s Experiment [22]

We need an existing user study's experiment data as an example to examine the generalization error rates and estimated type I/II errors. Thus, we chose to reproduce the experiment of an IIR study by Kelly et al. [22]. We examine the influence of the number of participants on generalization error rates using the experiment's study. We choose to reproduce this study because many later studies have used similar designs and tasks.

Kelly et al.'s study [22] has examined the influence of task cognitive complexity (independent variable) on search behavior and user experience measures (dependent variables). They designed ten tasks of five different cognitive complexity levels (Remember, Understand, Analyze, Evaluate, and Create). Their experiment used a within-subjects design and included 48 participants. Their study found that task complexity levels significantly affect many but not all search behavior and user experience measures.

We reproduced Kelly et al.'s study [22] in a crowdsourcing setting (instead of a lab-based one) due to the COVID-19 pandemic. Our reproducing experiment used the same independent and dependent variables, measurements for the variables (questions and items), and experiment stimuli (the ten tasks). But we made a few changes to the experimental design for the crowdsourcing platform. First, we used a between-subjects design, where each participant only needed to finish one task. Second, we set a minimum time duration a participant is required to spend on a task session. Third, our system returns results from the Bing search API while Kelly et al. let participants search the open web using the search engine of their choice. Table 2 summarizes the settings of the two experiments.

We required participants to spend at least 2, 4, 7, 8, and 9 minutes in tasks of the five task complexity levels (Remember, Understand, Analyze, Evaluate, and Create), respectively. This was to help the crowdsourcing participants maintain a similar level of engagement to those in Kelly et al.'s lab-based studies. In a pilot study, we let crowdsourcing participants search until they wanted to stop. However, they had spent a much shorter time in the search tasks than those reported by Kelly et al. After setting the minimum search time, participants' overall activities were as frequent as those in Kelly et al. reported.

3.4 Data and Estimation

We recruited 500 valid participants from Amazon Mechanical Turk for the formal study (100 for each IV level). Each participant needed to finish one task, including searching information and answering the pre-task and post-task questions. To improve the crowdsourcing participants' engagement, we informed the participants that 1) we will check their search activities to remove invalid HITs (e.g., just staying there to get paid), and 2) they will have the chance to receive a bonus if they complete the task perfectly.

We use the restricted estimation introduced in Section 3.1 to estimate generalization error rates by sampling $K = 100,000$ times. We use the methods introduced in Section 3.2 to estimate type I/II error rates (also sampling $K = 100,000$ times). We estimate DV's distributions using Gaussian distributions, where we estimate the population mean as the sample mean and population SD as the sample SD (with Bessel's correction).

4 RESULTS

In this section, we first compare our results with those of Kelly et al. [22] and report the consistent and inconsistent results of the two studies. Then we discuss the consistent and inconsistent variables separately. The discussion mainly contains the two Error Rates' patterns and interpretations. We also speculate the optimal sample size for a generalized study using the Error Rate's results. In the discussion, for the variables with consistent results from the two studies, we divide the variables into two parts according to whether they have a significant effect or not. For the variables with inconsistent results, we explore each variable's Error Rate separately for the cases with and without significant effect.

4.1 Our Results vs. Kelly et al. [22]'s Results

We calculated participants' search behaviors and answers to the pre-task and post-task questions in different cognitive complexity levels. Then we followed Kelly et al. [22] 's inspection method, using one-way ANOVA to test whether the independent variables have significant effects on each dependent variable ($\alpha = 0.01$).

Table 3 compares our online user study's results with Kelly et al. [22] 's lab user study results. Our significance tests' results on two-thirds of the dependent variables are consistent with Kelly et al. [22] 's, while one-third of the dependent variables' results are inconsistent with Kelly et al. [22] 's. The comparison results showed that our results had a high consistency with the lab study's [22] results despite many differences in experimental settings.

We found that four search behavior DVs did not show significant differences in our study but have significant differences in the lab user study. It may be because that the online crowd workers invested limited effort in the tasks, and they would easily give up if the cognitive demand were too high. Another possible reason is that the relatively high individual differences in our online studies.

We found that most participants' search experiences were significantly affected by task cognitive complexity levels both in the two studies, such as *Pre-Task: Q7*, *Pre-Task: Q10* and *Post-Task: Q10*. In contrast, some measures have different results between the online study and the lab study. Our study found no significant differences in task complexity (including *Pre-Task: Q4*, *Pre-Task: Q5* and *Pre-Task: Q6*) before completing the tasks. But these results are different from those of the lab study [22]. We also found significant differences in our study participants' post-task difficulty ratings and overall difficulty ratings. But in the lab user study [22], we only found significant effects on participants' expectation of difficulty in determining information's adequacy. Moreover, participants' perceived interest increase and engagement had no significant differences in our study but had a significant effect in the Kelly et al.'s lab study.

4.2 Dependent Variables with Significant Differences in Both Studies

In this part, we first analyze the Generalizability Error Rate and the Estimated Type II Error Rate of dependent variables that have significant effects both in our studies and Kelly et al. [22]'s study. We find that the two error rates are very close to 100% when n is small and then drop as the sample size becomes larger (See Figure 1). We then compare the change patterns between the two error rates and find that when the sample size is small, these variables' Estimated Type II Error Rate is always lower than their Generalizability Error Rate. It indicates that the estimates of type II errors will underestimate the actual error rate in IIR studies. While when the sample size gets larger, the similarity between the sample and the population increases, and the Generalizability Error rates are

Table 3: Consistent and Inconsistent Results of Our Study and Kelly et al. [22]'s study

Dependent Variables	Our Study	Kelly et al. [22]'s Study
UniqueQs (Number of unique queries submitted in a task)	Sig.	Sig.
QLen (Average number of query terms in all unique queries in a task)	No Sig.	Sig.
UniqueTerms (Number of unique query terms used in a task)	Sig.	Sig.
SERPClicks (Number of clicks made on SERPs in a task)	Sig.	Sig.
UniqueURLs (Number of unique URLs visited in a task)	Sig.	Sig.
QnoClick (Number of unique queries without clicks on SERP in a task)	Sig.	Sig.
QDiv (Number of queries that not issued by other participants completing the same task)	Sig.	Sig.
QTermDiv (Number of queries terms that not used by other participants completing the same task)	No Sig.	Sig.
URLDiv (Number of URLs that not visited by other participants completing the same task)	No Sig.	Sig.
Pre-Task: Q1 (Expectation of interest in the task)	No Sig.	No Sig.
Pre-Task: Q2 (Number of previous searches for task-related information)	No Sig.	No Sig.
Pre-Task: Q3 (Prior Knowledge of the task)	No Sig.	No Sig.
Pre-Task: Q4 (Expectation of types of information needed)	No Sig.	Sig.
Pre-Task: Q5 (Expectation of steps required to complete the task)	No Sig.	Sig.
Pre-Task: Q6 (Expectation of solutions)	No Sig.	Sig.
Pre-Task: Q7 (Expectation of difficulty in searching for information)	Sig.	Sig.
Pre-Task: Q8 (Expectation of difficulty in understanding information)	Sig.	Sig.
Pre-Task: Q9 (Expectation of difficulty in deciding information's usefulness)	Sig.	Sig.
Pre-Task: Q10 (Expectation of difficulty in integrating information)	Sig.	Sig.
Pre-Task: Q11 (Expectation of difficulty in determining information's adequacy)	Sig.	Sig.
Post-Task: Q1 (Perceived enjoyable)	No Sig.	No Sig.
Post-Task: Q2 (Perceived engagement)	No Sig.	Sig.
Post-Task: Q3 (Perceived difficulty in concentrating)	Sig.	No Sig.
Post-Task: Q4 (Perceived interest increase)	No Sig.	Sig.
Post-Task: Q5 (Perceived knowledge increase)	No Sig.	No Sig.
Post-Task: Q6 (Perceived difficulty in searching for information)	No Sig.	No Sig.
Post-Task: Q7 (Perceived difficulty in understanding information)	No Sig.	No Sig.
Post-Task: Q8 (Perceived difficulty in deciding information's usefulness)	No Sig.	No Sig.
Post-Task: Q9 (Perceived difficulty in integrating information)	Sig.	No Sig.
Post-Task: Q10 (Perceived difficulty in determining information's adequacy)	Sig.	Sig.
Post-Task: Q11 (Perceived overall difficulty)	Sig.	No Sig.
Post-Task: Q12 (Perceived satisfaction with the solution)	No Sig.	No Sig.
Post-Task: Q13 (Perceived satisfaction with search strategy)	No Sig.	No Sig.

underestimated to a greater extent. Therefore, when the sample size is large, we found the Estimated Type II Error Rate is higher than the Generalizability Error Rate.

We record the minimum sample size when the two error rates drop below 20% and 10% (See Table 4). Several dependent variables' error rates drop at a rapid pace as the sample size becomes larger. Therefore, their Generalizability Error Rate can decline to 20% and even 10% when the sample size is relatively small. The Estimated Type II Error Rate of these variables also decreased quickly and arrived at 20% and 10% when their simulated sample size was similar to that of the Generalizability Error Rate. Take *QDiv* as an example (See Figure 2(a)). Its Generalizability Error Rate is higher than its Estimated Type II Error Rate when the sample size is smaller than 20. Then the Estimated Type II Error Rate exceeds the Generalizability Error Rate until the sample size reaches 70, and finally, the two error rates both reach 0%. Because the Generalizability Error Rate decreases faster than the Estimated Type II Error Rate, it reaches the 20% and 10% thresholds when its sample size is 30 and 35. In comparison, the Estimated Type II Error Rate cannot reach 20% and 10% until its sample size is 35 and 45.

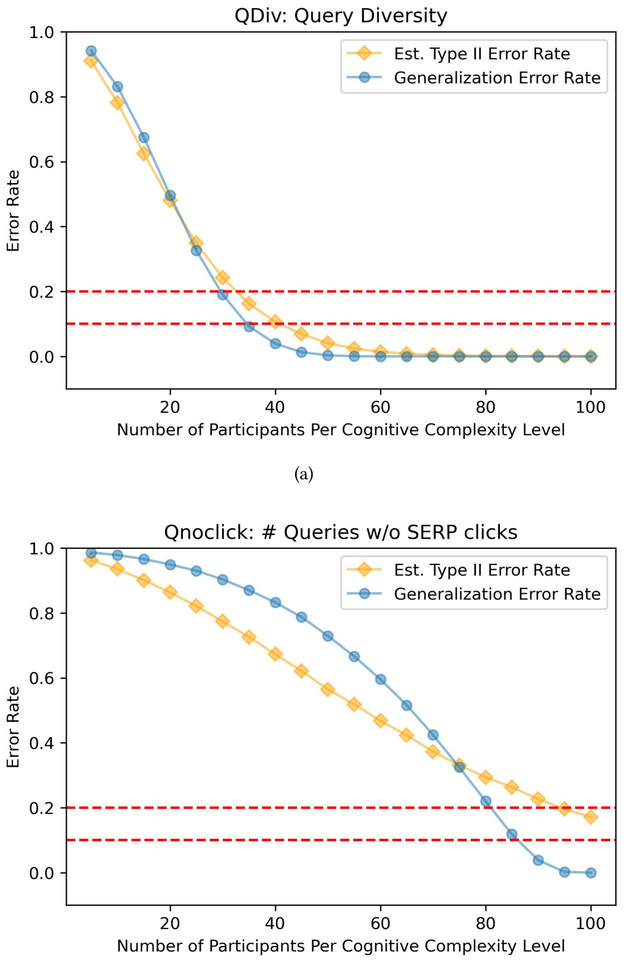
We also find that some variables' Generalizability Error Rates decrease slower than the previous cases, and the sample size required to reach the thresholds is larger than that in the previous cases (See Figure 2(b)). For example, *QnoClick*'s Generalizability Error Rate declined slowly and maintained a leading edge in the Estimated Type II Error Rate before the sample size reached 75. Then its Generalizability Error Rate turns to be lower than its Estimated Type II Error Rate when the sample size is larger than 75. Its Generalizability Error Rate reaches 20% when the sample size is 80 and reaches 10% when the sample size is 60. Moreover, *QnoClick*'s Estimated Type II Error Rate cannot reach all thresholds when the sample size is smaller than 100: it reaches 20% when the sample size is 95 and reaches 10% when the sample size is 120.

Table 4: Minimum Sample Size for Dependent Variables' Generalization Error Rate and Gaussian Distribution Error Rate to Reach Beta Thresholds

	Generalization Error Rate		Estimated Type II Error Rate	
DV	beta<20%	beta<10%	beta<20%	beta<10%
UniqueTerms*	5	5	5	5
QDiv*	30	35	35	45
UniqueQs*	40	45	40	50
Post-Task: Q11*	50	60	50	65
Pre-Task: Q7*	60	70	65	75
Post-Task: Q10*	65	75	70	85
Pre-Task: Q11*	75	80	75	90
SERPClicks*	75	80	80	100
UniqueURLs*	75	80	80	100
Pre-Task: Q10*	85	90	90	120
Post-Task: Q3*	85	90	90	110
QnoClick*	85	90	95	120
Pre-Task: Q8*	90	95	110	130
Post-Task: Q9*	90	95	100	120
Pre-Task: Q9*	/	/	120	150
URLDiv	/	/	130	150
Pre-Task: Q4	/	/	140	180
Post-Task: Q7	/	/	130	160
Post-Task: Q4	/	/	150	190
Pre-Task: Q1	/	/	360	440
Pre-Task: Q2	/	/	370	460
Pre-Task: Q3	/	/	160	200
Pre-Task: Q5	/	/	430	/
Pre-Task: Q6	/	/	170	220
Post-Task: Q1	/	/	400	490
Post-Task: Q2	/	/	300	370
Post-Task: Q5	/	/	190	230
Post-Task: Q6	/	/	170	210
Post-Task: Q8	/	/	160	190
Post-Task: Q12	/	/	300	370
Post-Task: Q13	/	/	210	260
Qlen	/	/	160	190
QTermDiv	/	/	420	/

* represents the variable is significant in our study, and the light blue cell represents the variable is significant in Kelly et al. [22]'s study

Figure 1: Error Rates of Dependent Variables with Significant Differences in Both Studies



(a)

(b)

4.3 Dependent Variables without Significant Differences in Both Studies

When independent variables had no significant effects both in our studies and Kelly et al. [22]'s study, the change patterns of their Generalizability Error Rate are very different. We record the minimum sample size when the Generalizability Error Rate drops to certain thresholds (See Table 5). Because the Estimated type I Error Rate of variables without significant difference is the same as the significance level (alpha = 0.01), the Estimated type I Error Rate is very close to 0.01 in this part.

There are several dependent variables' error rate curves first rise and then fall, but their change rates were different (See Figure 2). The increase of the Generalizability Error Rate at the beginning is because we put the results of dependent variables without significant effects as the basic standard of the Generalizability Error Rate's calculation. Therefore, the more significant effect sample data has, the higher the error rate is. When the sample size is small, it is hard to detect the differences, so the low error rate. When the sample size gets larger, the possibility of extreme values in the data will increase so that we can find more significant differences. However, because the sample population is known, the similarity between the population and the sample increases as the sample size gets larger. Therefore, the error rate will become stable and decline when the sample size approaches the population. Moreover, different change rates may be caused by the uneven data distribution: if there are many extreme values in the data, the results are more likely to be significant.

Although these variables' error rate has a rise and fall change pattern, their Generalizability Error Rate remains relatively low. Take *Pre-Task: Q3* as an example (See Figure 3(a)). Its Generalizability Error Rate is nearly 2% when the sample size is 5. Then it

Table 5: Minimum Sample Size for Dependent Variables' Generalization Error Rate to Reach Alpha Thresholds

	Generalization Error Rate		Generalization Error Rate
DV	alpha=0.01	DV	alpha=0.01
QTermDiv	5	Pre-Task: Q11*	/
Pre-Task: Q2	10	Post-Task: Q3*	/
Post-Task: Q1	10	Post-Task: Q4	/
Pre-Task: Q5	25	Post-Task: Q7	/
Pre-Task: Q1	30	Post-Task: Q8	/
Post-Task: Q2	45	Post-Task: Q9*	/
Post-Task: Q12	65	Post-Task: Q10*	/
Post-Task: Q13	80	Post-Task: Q11*	/
Post-Task: Q5	95	UniqueQs*	/
Post-Task: Q6	95	Qlen	/
Pre-Task: Q3	/	UniqueTerms*	/
Pre-Task: Q4	/	SERPClicks*	/
Pre-Task: Q6	/	UniqueURLs*	/
Pre-Task: Q7*	/	QnoClick*	/
Pre-Task: Q8*	/	QDiv*	/
Pre-Task: Q9*	/	URLDiv	/
Pre-Task: Q10*	/		

* represents the variable is significant in our study, and the light blue cell represents the variable is significant in Kelly et al. [22]'s study

begins to rise at a rapid rate to nearly 12%. After that, the curve becomes stable and starts to decrease. It decreased under 10% when the sample size was 95. We can conclude that users have diverse satisfaction with their search strategies so that the differences of small sample size are significant. We can also find that the sample size of 100 does not meet the need of conducting a study of high generalizability. The Generalizability Error Rate will decrease to 5% and 0% smoothly when its population is large enough.

Another example of these variables is *Post-Task: Q12* (See Figure 3(b)), its error rate rises and falls at a slower rate. Because its error rate drops after a slight increase, its curve's starting point and the highest point are both around 2%. Its Generalizability Error Rate decreased under 1% threshold when the sample size is 60 and reached 0% when the sample size is 90. Since the sample size of 90 is very close to the sample size of 100, we need to increase the number of participants in future work to remove the influence of similarity with the original data.

Some variables' error rate curves first rise and do not fall until the sample size is very large. For example, *Post-Task: Q7* 's error rate is around 1.5% when the sample size is 5. Then it keeps increasing and reaches nearly 0.2 when the sample size is 90 (See Figure 3(c)). Its error rate suddenly drops to 0 when the sample size is 100. This result indicates that the sample data always has a significant effect regardless of the sample size, although the population has no significant effect. We set the ANOVA test's significance level at 0.01, and these variables' populations have no significant effect. While when we set the significance level at 0.05, they do have a significant effect like all the sample data. Therefore, it reminds us that when testing whether a variable has a significant effect, we should not restrict the significance level too tightly, especially since the variables' p-value is slightly larger than 0.01 and smaller than 0.05.

In addition to the change patterns above, some variables also have a Generalizability Error Rate curve like the first case, but without a small increase before the decline. *Post-Task: Q1* belongs to this kind of variables (See Figure 3(d)). Its curve decreased under 1% when the sample size was nearly 15 and reached 0% when it was 80. The reason for the disappearance of the small increase may be that the data distribution is very even so that extreme values' influence is very small. Therefore, the data do not have more significant differences when the sample size gets larger.

4.4 Dependent Variables had Inconsistent Results in Two Studies

As mentioned in 4.1, the results of one-third of independent variables in our experiment are inconsistent with the results of Kelly et al. [22] 's. Some variables have significant effects in our study but have no significant effect in Kelly et al. [22] 's, and some other variables have no significant effect in our experiment but have a significant effect in Kelly et al. [22] 's.

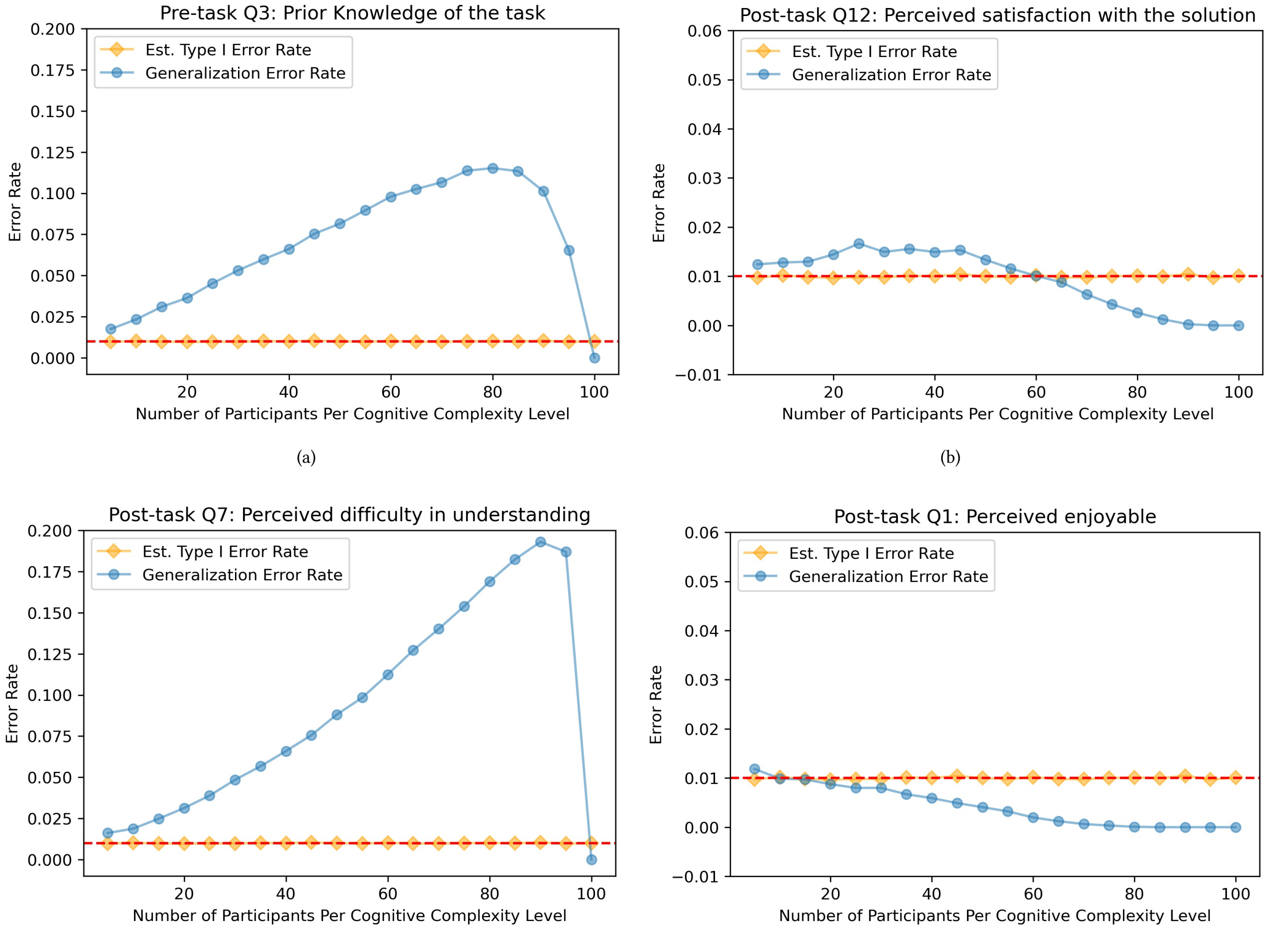
These variables all have two possible results: significant and not significant, and we have different analysis methods for these two kinds of results. Therefore, we will explore the change patterns of the Generalizability Error Rate under the two results separately. If all independent variables with inconsistent results in the two studies do have significant results, their Generalizability Error Rates have different change patterns (See Figure 3).

We first explore cases that all the variables have significant effects. There are several independent variables' Generalizability Error Rate drops at a very slow rate and cannot even reach the 20% threshold before the sample size reach 100. While it suddenly drops to 0% when the sample size reaches 100 (See Figure 4(a)). *Post-Task: Q9* is one of these kinds of variables. Its Generalizability Error Rate declines very slow and is always higher than the Estimated type II Error Rate before the sample size reaches 85. When the sample size is higher than 85, its Generalizability Error Rate turns to be lower than its Estimated type II Error Rate. Its Generalizability Error Rate reaches 20% threshold when the sample size is 90 and reaches 10% threshold when the sample size is 95. In contrast, its Estimated type II Error Rate reaches 20% when the sample size is 100 and reaches 10% when the sample size is 120. It indicated that 100 participants for each task's cognitive complexity level were insufficient for detecting these variable's effects. Studies related to these variables and want to make sure to have generalizability should recruit more than 100 participants.

Some of these change patterns have been explained in 4.2. For example, *Post-Task: Q11* 's Generalizability Error Rate drop at a rapid rate as the sample size becomes larger (See Figure 4(b)). Its Generalizability Error Rate decreases to 20% threshold when the sample size is 50 and reaches 10% threshold when the sample size is 60. Moreover, its Estimated type II Error Rate can also reach all thresholds when the sample size is slightly larger but still smaller than 100. Its change pattern is similar to the first case in 4.2.

There is also some variables' Generalizability Error Rates first decrease keeps slowly from 100% and then rising back to 100%. When the sample size reaches 100, their Generalizability Error

Figure 2: Error Rates of Dependent Variables without Significant Difference in Both Studies



(a)

(b)

(c)

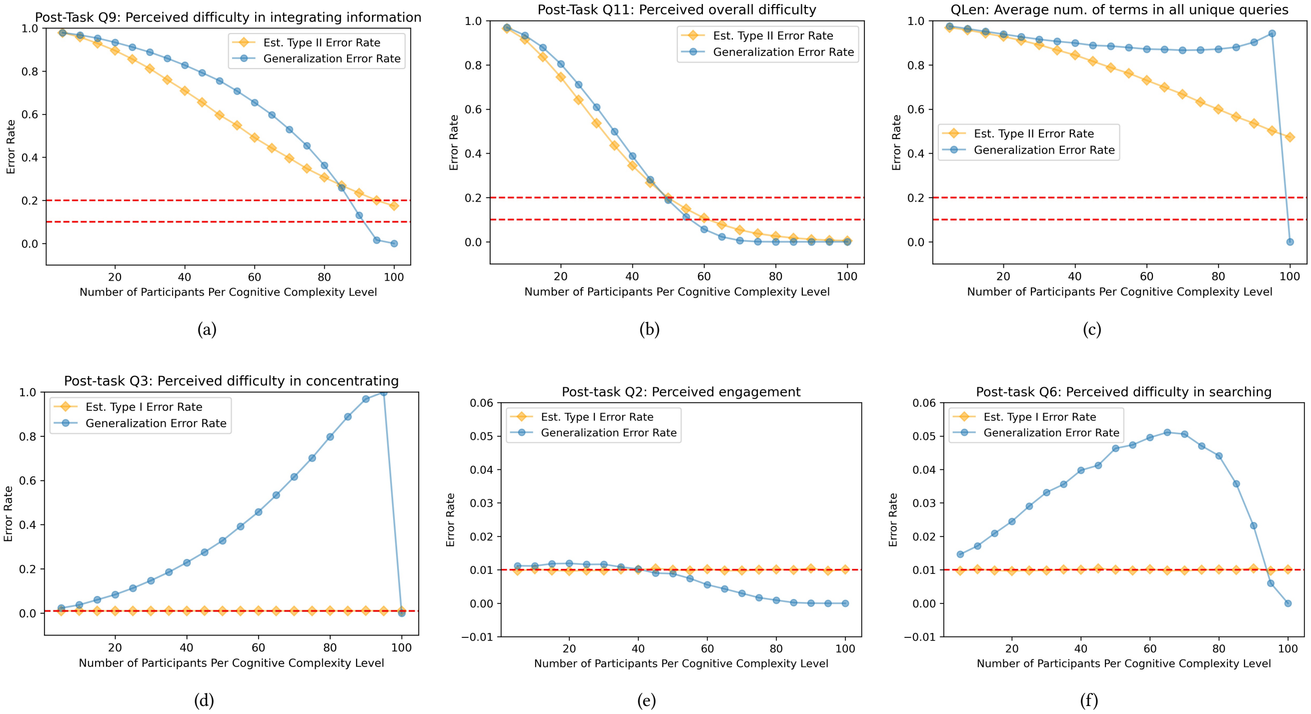
(d)

Rates suddenly drop to 0%. At the same time, their Estimated type II Error Rate keeps declining but cannot reach any thresholds when the sample size is smaller than 100 (See Figure 4(c)). For example, *QLen* 's Generalizability Error Rate begins at nearly 95% when the sample size is 5, then first declines and then increases as the sample size gets larger. Its Estimated type II Error Rate keeps dropping and reaches 20% when the sample size is 160, then reaches 10% when the sample size is smaller than 190.

Then we start to explore cases that the variables that have no significant effects. The data of our population is insignificant, but we still can get several implications from the analysis. As mentioned in 4.3, the Estimated type I Error Rate of insignificant variables is the same as the significance level, and thus this error rate keeps being 0.01 in this part. Some of these variables' Generalizability Error Rate keep rising and do not fall until the sample size is 100. For example, *Post-Task: Q3* 's Generalizability Error Rate was 2% when the sample size is 5, then it keeps increasing and reaches nearly 100% when the sample size is 95 (See Figure 4(d)). When the sample size is 100, its error rate suddenly drops to 0%. This changing pattern of Generalizability Error Rate has been discussed in 4.3. We can conclude that the population has no significant effect, but the sample data always has a significant effect.

For the variables that are insignificant regardless of the significance level, their Generalizability Error Rates' curve first rise and then fall to nearly 0% (See Figure 4(e)). *Post-Task: Q2* is one of these variables. It begins at nearly 1%, then decreases after a slight increase. As explained in 4.3, the increase of Generalizability Error Rate at the beginning is because when the sample size gets larger, more significant differences will be found as the number of extreme values increases. The following decrease in Generalizability Error Rate is due to the similarity between the population and the sample increases as the sample size gets larger. This pattern is more obvious in some other variables. For example, *Pre-Task: Q6* begins at 1.5% when its sample size is 5. It keeps increasing and reaches it highest point at 8% when its sample size is nearly 70, then it becomes stable and begins to drop off. Its Generalizability Error Rate cannot decline to 1% until the sample size is 100. It indicates that larger sample size is needed to find the minimum number of participants of a generalized study.

Figure 3: Error Rates of Dependent Variables had Inconsistent Results in Two Studies



(a)

(b)

(c)

(d)

(e)

(f)

5 DISCUSSION

We have performed an analysis regarding how the number of participants in IIR user studies influences the experiment’s findings (determined by significance tests). We have examined an experiment that successfully reproduced a popular IIR study using task complexity levels as independent variables.

Our study contributes to the current understandings from two different aspects:

First, we have compared generalization error rates based on samples drawn from actual observations with estimated type I and II errors based on data distribution assumptions. The latter is equivalent to the analysis of type I and II errors in statistics based on the alpha level and power analysis. However, it was worth noting that the generalization error rates at a small sample size are usually higher than the estimated type I and II errors. Also, as we discussed in Section 3, our estimation underestimates the generalization error rates. Thus, the estimated values based on random samples drawn from the total participants only provide a lowerbound of the actual value. This is to say that the actual generalization error rates will be even higher than the estimated ones and type I/II errors. It is clear that the estimate of type I and II errors based on data distribution assumptions (as most significance tests require) is largely underestimated. The extent to which it was underestimated seems to vary by variable, suggesting that in practice, we may expect 1) significance tests to have higher than theoretical type I and II error rates and 2) have different actual type I/II error rates when performing on different variables.

Second, we have chosen an influential study by Kelly et al. for a case study, and our results facilitate future experiments using a similar design. Kelly et al.’s study provided an infrastructure for designing search tasks based on cognitive complexity levels. The ten tasks they used have been approved (by both their study and our work) to be successful for their purpose (providing stimuli search tasks of varied task complexity levels). Our successful reproduction of Kelly et al.’s study in a different experiment setting further verified the correctness of their most findings. More importantly, we further suggested the number of participants needed to reduce the generalization error rates or estimated type I/II error rates to a certain level. This can largely benefit future studies using a similar experimental design.

Our work is also limited in several places. First, we have employed a limited number of participants for our purpose. Although we have used twice amount of participants as most other IIR user studies did, it seems that the sample size is still not enough when estimating error rates for certain variables. This influences the accuracy of some findings, e.g., the suggested number of participants. Second, we have used online experiments instead of lab-based ones (the dominating approach in IIR research) due to the COVID-19 pandemic. This restricts the usefulness of the specific findings (e.g., the suggested number of participants) for future lab-based experiments. Also, we expect future work to improve our study by including more experiments for analysis.

6 CONCLUSION

To conclude, our study has examined the influence of the number of participants on a specific IIR user study reproducing a popular one by Kelly et al. However, our method for examining this matter, the generalization error rate measure, can be applied to analyze other IIR experiments too. The findings have disclosed that the theoretical estimates of type I and II errors (such as the alpha level and the power analysis) are underestimated. Also, we offer suggestions regarding the number of participants for future studies using task complexity levels as independent variables. Our work provides a guideline for IIR user studies in the future.

ACKNOWLEDGEMENT

Support for this research was provided by anonymous funding agency. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

REFERENCES

[1] Ioannis Arapakis, Xiao Bai, and B. Barla Cambazoglu. 2014. Impact of Response Latency on User Behavior in Web Search. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval* (Gold Coast, Queensland, Australia) *(SIGIR '14)*. Association for Computing Machinery, New York, NY, USA, 103–112. https://doi.org/10.1145/2600428.2609627

[2] Sandeep Avula, Gordon Chadwick, Jaime Arguello, and Robert Capra. 2018. SearchBots: User Engagement with ChatBots during Collaborative Search. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (New Brunswick, NJ, USA) *(CHIIR '18)*. Association for Computing Machinery, New York, NY, USA, 52–61. https://doi.org/10.1145/3176349.3176380

[3] Leif Azzopardi. 2014. Modelling Interaction with Economic Models of Search. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval* (Gold Coast, Queensland, Australia) *(SIGIR '14)*. Association for Computing Machinery, New York, NY, USA, 3–12. https://doi.org/10.1145/2600428.2609574

[4] Miguel Barreda-Ángeles, Ioannis Arapakis, Xiao Bai, B. Barla Cambazoglu, and Alexandre Pereda-Baños. 2015. Unconscious Physiological Effects of Search Latency on Users and Their Click Behaviour. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Santiago, Chile) *(SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 203–212. https://doi.org/10.1145/2766462.2767719

[5] Andrei Broder. 2002. A Taxonomy of Web Search. *SIGIR Forum* 36, 2 (Sept. 2002), 3–10. https://doi.org/10.1145/792550.792552

[6] Chris Buckley and Ellen M. Voorhees. 2000. Evaluating Evaluation Measure Stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Athens, Greece) *(SIGIR '00)*. Association for Computing Machinery, New York, NY, USA, 33–40. https://doi.org/10.1145/345508.345543

[7] Ryan Burton and Kevyn Collins-Thompson. 2016. User Behavior in Asynchronous Slow Search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 345–354. https://doi.org/10.1145/2911451.2911541

[8] Robert Capra, Jaime Arguello, Anita Crescenzi, and Emily Vardell. 2015. Differences in the Use of Search Assistance for Tasks of Varying Complexity. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Santiago, Chile) *(SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 23–32. https://doi.org/10.1145/2766462.2767741

[9] Jesse David Dinneen, Banafsheh Asadi, Ilja Frissen, Fei Shu, and Charles-Antoine Julien. 2018. Improving Exploration of Topic Hierarchies: Comparative Testing of Simplified Library of Congress Subject Heading Structures. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (New Brunswick, NJ, USA) *(CHIIR '18)*. Association for Computing Machinery, New York, NY, USA, 102–109. https://doi.org/10.1145/3176349.3176385

[10] Ashlee Edwards and Diane Kelly. 2017. Engaged or Frustrated? Disambiguating Emotional State in Search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 125–134. https://doi.org/10.1145/3077136.3080818

[11] Jean Garcia-Gathright, Brian St. Thomas, Christine Hosey, Zahra Nazari, and Fernando Diaz. 2018. Understanding and Evaluating User Satisfaction with Music Discovery. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 55–64. https://doi.org/10.1145/3209978.3210049

[12] Souvick Ghosh, Manasa Rath, and Chirag Shah. 2018. Searching as Learning: Exploring Search Behavior and Learning Outcomes in Learning-Related Tasks. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (New Brunswick, NJ, USA) *(CHIIR '18)*. Association for Computing Machinery, New York, NY, USA, 22–31. https://doi.org/10.1145/3176349.3176386

[13] Morgan Harvey, Claudia Hauff, and David Elsweiler. 2015. Learning by Example: Training Users with High-Quality Query Suggestions. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Santiago, Chile) *(SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 133–142. https://doi.org/10.1145/2766462.2767731

[14] Morgan Harvey and Matthew Pointon. 2017. Searching on the Go: The Effects of Fragmented Attention on Mobile Web Search Tasks. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 155–164. https://doi.org/10.1145/3077136.3080770

[15] Jiyin He, Marc Bron, Arjen de Vries, Leif Azzopardi, and Maarten de Rijke. 2015. Untangling Result List Refinement and Ranking Quality: A Framework for Evaluation and Prediction. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Santiago, Chile) *(SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 293–302. https://doi.org/10.1145/2766462.2767740

[16] Jiyin He and Emine Yilmaz. 2017. User Behaviour and Task Characteristics: A Field Study of Daily Information Behaviour. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (Oslo, Norway) *(CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 67–76. https://doi.org/10.1145/3020165.3020188

[17] Daniel Hienert, Matthew Mitsui, Philipp Mayr, Chirag Shah, and Nicholas J. Belkin. 2018. The Role of the Task Topic in Web Search of Different Task Types. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (New Brunswick, NJ, USA) *(CHIIR '18)*. Association for Computing Machinery, New York, NY, USA, 72–81. https://doi.org/10.1145/3176349.3176382

[18] Orland Hoeber, Anoop Sarkar, Andrei Vacariu, Max Whitney, Manali Gaikwad, and Gursimran Kaur. 2017. Evaluating the Value of Lensing Wikipedia During the Information Seeking Process. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (Oslo, Norway) *(CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 77–86. https://doi.org/10.1145/3020165.3020178

[19] Jiacheng Huang, Wei Hu, Haoxuan Li, and Yuzhong Qu. 2018. Automated Comparative Table Generation for Facilitating Human Intervention in Multi-Entity Resolution. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 585–594. https://doi.org/10.1145/3209978.3210021

[20] Jiepu Jiang, Daqing He, and James Allan. 2017. Comparing In Situ and Multidimensional Relevance Judgments. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 405–414. https://doi.org/10.1145/3077136.3080840

[21] Jiepu Jiang, Daqing He, Diane Kelly, and James Allan. 2017. Understanding Ephemeral State of Relevance. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (Oslo, Norway) *(CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 137–146. https://doi.org/10.1145/3020165.3020176

[22] Diane Kelly, Jaime Arguello, Ashlee Edwards, and Wan-ching Wu. 2015. Development and Evaluation of Search Tasks for IIR Experiments Using a Cognitive Complexity Framework. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval* (Northampton, Massachusetts, USA) *(ICTIR '15)*. Association for Computing Machinery, New York, NY, USA, 101–110. https://doi.org/10.1145/2808194.2809465

[23] Jaewon Kim, Paul Thomas, Ramesh Sankaranarayana, Tom Gedeon, and Hwan-Jin Yoon. 2017. What Snippet Size is Needed in Mobile Web Search?. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (Oslo, Norway) *(CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 97–106. https://doi.org/10.1145/3020165.3020173

[24] Jin Young Kim, Nick Craswell, Susan Dumais, Filip Radlinski, and Fang Liu. 2017. Understanding and Modeling Success in Email Search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 265–274. https://doi.org/10.1145/3077136.3080837

[25] Julia Kiseleva, Kyle Williams, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Predicting User Satisfaction with Intelligent Assistants. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 45–54. https://doi.org/10.1145/2911451.2911521

[26] Khalil Klouche, Tuukka Ruotsalo, Luana Micallef, Salvatore Andolina, and Giulio Jacucci. 2017. Visual Re-Ranking for Multi-Aspect Information Retrieval. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (Oslo, Norway) *(CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 57–66. https://doi.org/10.1145/3020165.3020174

[27] Michael Kotzyba, Tatiana Gossen, Johannes Schwerdt, and Andreas Nürnberger. 2017. Exploration or Fact-Finding: Inferring User's Search Activity Just in Time.

In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (Oslo, Norway) *(CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 87–96. https://doi.org/10.1145/3020165.3020180

[28] Yuelin Li and Nicholas J. Belkin. 2008. A Faceted Approach to Conceptualizing Tasks in Information Seeking. *Inf. Process. Manage.* 44, 6 (Nov. 2008), 1822–1837. https://doi.org/10.1016/j.ipm.2008.07.005

[29] Chenjun Ling, Ben Steichen, and Alexander G. Choulos. 2018. A Comparative User Study of Interactive Multilingual Search Interfaces. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (New Brunswick, NJ, USA) *(CHIIR '18)*. Association for Computing Machinery, New York, NY, USA, 211–220. https://doi.org/10.1145/3176349.3176383

[30] Xiaozhong Liu, Zhuoren Jiang, and Liangcai Gao. 2015. Scientific Information Understanding via Open Educational Resources (OER). In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Santiago, Chile) *(SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 645–654. https://doi.org/10.1145/2766462.2767750

[31] Yiqun Liu, Ye Chen, Jinhui Tang, Jiashen Sun, Min Zhang, Shaoping Ma, and Xuan Zhu. 2015. Different Users, Different Opinions: Predicting Search Satisfaction with Mouse Movement Information. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Santiago, Chile) *(SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 493–502. https://doi.org/10.1145/2766462.2767721

[32] Yiqun Liu, Zeyang Liu, Ke Zhou, Meng Wang, Huanbo Luan, Chao Wang, Min Zhang, and Shaoping Ma. 2016. Predicting Search User Examination with Visual Saliency. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 619–628. https://doi.org/10.1145/2911451.2911517

[33] Zeyang Liu, Yiqun Liu, Ke Zhou, Min Zhang, and Shaoping Ma. 2015. Influence of Vertical Result in Web Search Examination. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Santiago, Chile) *(SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 193–202. https://doi.org/10.1145/2766462.2767714

[34] Hongyu Lu, Min Zhang, and Shaoping Ma. 2018. Between Clicks and Satisfaction: Study on Multi-Phase User Preferences and Satisfaction for Online News Reading. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 435–444. https://doi.org/10.1145/3209978.3210007

[35] Cheng Luo, Xue Li, Yiqun Liu, Tetsuya Sakai, Fan Zhang, Min Zhang, and Shaoping Ma. 2017. Investigating Users' Time Perception during Web Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (Oslo, Norway) *(CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 127–136. https://doi.org/10.1145/3020165.3020184

[36] Cheng Luo, Yiqun Liu, Tetsuya Sakai, Fan Zhang, Min Zhang, and Shaoping Ma. 2017. Evaluating Mobile Search with Height-Biased Gain. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 435–444. https://doi.org/10.1145/3077136.3080795

[37] David Maxwell, Leif Azzopardi, and Yashar Moshfeghi. 2017. A Study of Snippet Length and Informativeness: Behaviour, Performance and User Experience. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 135–144. https://doi.org/10.1145/3077136.3080824

[38] Florian Meier and David Elsweiler. 2016. Going Back in Time: An Investigation of Social Media Re-Finding. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 355–364. https://doi.org/10.1145/2911451.2911524

[39] Yashar Moshfeghi, Peter Triantafillou, and Frank E. Pollick. 2016. Understanding Information Need: An FMRI Study. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 335–344. https://doi.org/10.1145/2911451.2911534

[40] Kevin Ong, Kalervo Järvelin, Mark Sanderson, and Falk Scholer. 2017. Using Information Scent to Understand Mobile and Desktop Web Search Behavior. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 295–304. https://doi.org/10.1145/3077136.3080817

[41] Tetsuya Sakai. 2016. Statistical Significance, Power, and Sample Sizes: A Systematic Review of SIGIR and TOIS, 2006-2015. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 5–14. https://doi.org/10.1145/2911451.2911492

[42] Joni Salminen, Bernard J. Jansen, Jisun An, Soon-Gyo Jung, Lene Nielsen, and Haewoon Kwak. 2018. Fixation and Confusion: Investigating Eye-Tracking Participants' Exposure to Information in Personas. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (New Brunswick, NJ, USA) *(CHIIR '18)*. Association for Computing Machinery, New York, NY, USA, 110–119. https://doi.org/10.1145/3176349.3176391

[43] Mark Sanderson and Justin Zobel. 2005. Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Salvador, Brazil) *(SIGIR '05)*. Association for Computing Machinery, New York, NY, USA, 162–169. https://doi.org/10.1145/1076034.1076064

[44] Bahareh Sarrafzadeh and Edward Lank. 2017. Improving Exploratory Search Experience through Hierarchical Knowledge Graphs. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 145–154. https://doi.org/10.1145/3077136.3080829

[45] Jaspreet Singh, Sergej Zerr, and Stefan Siersdorfer. 2017. Structure-Aware Visualization of Text Corpora. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (Oslo, Norway) *(CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 107–116. https://doi.org/10.1145/3020165.3020182

[46] Yu Su, Ahmed Hassan Awadallah, Miaosen Wang, and Ryen W. White. 2018. Natural Language Interfaces with Fine-Grained User Interaction: A Case Study on Web APIs. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 855–864. https://doi.org/10.1145/3209978.3210013

[47] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the Design of Spoken Conversational Search: Perspective Paper. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (New Brunswick, NJ, USA) *(CHIIR '18)*. Association for Computing Machinery, New York, NY, USA, 32–41. https://doi.org/10.1145/3176349.3176387

[48] Andrew Turpin, Falk Scholer, Stefano Mizzaro, and Eddy Maddalena. 2015. The Benefits of Magnitude Estimation Relevance Assessments for Information Retrieval Evaluation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Santiago, Chile) *(SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 565–574. https://doi.org/10.1145/2766462.2767760

[49] Kazutoshi Umemoto, Takehiro Yamamoto, and Katsumi Tanaka. 2016. ScentBar: A Query Suggestion Interface Visualizing the Amount of Missed Relevant Information for Intrinsically Diverse Search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 405–414. https://doi.org/10.1145/2911451.2911546

[50] Ellen M. Voorhees and Chris Buckley. 2002. The Effect of Topic Set Size on Retrieval Experiment Error. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Tampere, Finland) *(SIGIR '02)*. Association for Computing Machinery, New York, NY, USA, 316–323. https://doi.org/10.1145/564376.564432

[51] Jieyu Wang and Anita Komlodi. 2018. Switching Languages in Online Searching: A Qualitative Study of Web Users' Code-Switching Search Behaviors. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (New Brunswick, NJ, USA) *(CHIIR '18)*. Association for Computing Machinery, New York, NY, USA, 201–210. https://doi.org/10.1145/3176349.3176396

[52] Yiwei Wang, Shawon Sarkar, and Chirag Shah. 2018. Juggling with Information Sources, Task Type, and Information Quality. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (New Brunswick, NJ, USA) *(CHIIR '18)*. Association for Computing Machinery, New York, NY, USA, 82–91. https://doi.org/10.1145/3176349.3176390

[53] Wan-Ching Wu, Diane Kelly, Ashlee Edwards, and Jaime Arguello. 2012. Grannies, Tanning Beds, Tattoos and NASCAR: Evaluation of Search Tasks with Varying Levels of Cognitive Complexity. In *Proceedings of the 4th Information Interaction in Context Symposium* (Nijmegen, The Netherlands) *(IIIX '12)*. Association for Computing Machinery, New York, NY, USA, 254–257. https://doi.org/10.1145/2362724.2362768

[54] Xiaohui Xie, Yiqun Liu, Xiaochuan Wang, Meng Wang, Zhijing Wu, Yingying Wu, Min Zhang, and Shaoping Ma. 2017. Investigating Examination Behavior of Image Search Users. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 275–284. https://doi.org/10.1145/3077136.3080799

[55] Fan Zhang, Ke Zhou, Yunqiu Shao, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. How Well Do Offline and Online Evaluation Metrics Measure User Satisfaction in Web Image Search?. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 615–624. https://doi.org/10.1145/3209978.3210059

[56] Haotian Zhang, Mustafa Abualsaud, and Mark D. Smucker. 2018. A Study of Immediate Requery Behavior in Search. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (New Brunswick, NJ, USA) *(CHIIR '18)*. Association for Computing Machinery, New York, NY, USA, 181–190. https://doi.org/10.1145/3176349.3176400