Using Principal Component Analysis to Better Understand Behavioral Measures and their Effects

Jaime Arguello and Anita Crescenzi
School of Information and Library Science
University of North Carolina at Chapel Hill
[jarguello,amcc]@email.unc.edu

ABSTRACT

An important question in interactive IR research is: What do search behaviors tell us about specific task characteristics, post-task perceptions, and post-task outcomes such as knowledge gains? Much research has explored this question from different perspectives. A common approach is to consider a *wide* range of behavioral measures and examine their differences based on dependent variables of interest (e.g., post-task perceptions). In this paper, we use principal component analysis (PCA), a dimensionality reduction technique, to analyze behavioral measures captured during three previously published studies. Using PCA, we examine the underlying phenomena being captured by different behavioral measures, and we examine the influence of these phenomena on different outcomes related to participants' post-task perceptions (e.g., workload, difficulty, engagement, etc.). We argue (and show) that PCA can provide several benefits. First, it can help us understand the behavioral phenomena captured by different measures. Second, it can help us determine which measures are ambiguous or unambiguous with respect to the underlying phenomena being captured. Third, it can help us understand how behavioral phenomena (vs. individual measures) relate to searchers' perceptions of their search experience.

ACM Reference Format:
Jaime Arguello and Anita Crescenzi. 2019. Using Principal Component Analysis to Better Understand Behavioral Measures and their Effects. In *The 2019 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '19), October 2–5, 2019, Santa Clara, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3341981.3344222

1 INTRODUCTION

An important question in interactive information retrieval (IIR) research is: What do search behaviors tell us about specific characteristics of a user's task and/or a user's experience while searching for information? Much research has explored this question from different perspectives. In terms of task characteristics, prior work has aimed at understanding how behavioral measures are influenced by task characteristics such as the task *type* (e.g., exploratory vs. known-item) [3], task *goal* (e.g., specific vs. amorphous) [18, 22, 24], task *product* (e.g., factual vs. intellectual) [18, 22, 24], and task *complexity* [20]. In terms of post-task perceptions, studies have considered how search behaviors are correlated with perceptions of task difficulty [1, 4, 9, 21, 23], frustration [11, 13], time pressure [8, 9] and engagement [11, 27]. Finally, in recent years, studies have examined how search behaviors can predict knowledge gains [7, 12, 14, 26]. Understanding the relation between behavioral measures and task characteristics can inform the design of dynamic support tools. Similarly, understanding the relation between behavioral measures and post-task perceptions and outcomes can help us evaluate systems using behavioral data alone.

In the studies mentioned above, a common approach is to generate a *wide* range of behavioral measures and examine how these vary based on task characteristics or post-task perceptions/outcomes. To illustrate, Jiang et al. [18] compared search behaviors across four task types: known-item, known-subject, interpretive, and exploratory. The authors considered 44 measures derived from SERP-level interactions (e.g., % abandoned queries), eye-tracking data (e.g., avg. rank examined), and query-reformulation behavior (e.g., avg. result overlap between queries). Zhuang et al [27] studied the relation between 37 measures and the levels of engagement reported by participants after a task. Yu et al. [26] examined the relation between 40 behavioral measures and participants' knowledge gains. In our own prior work, we used 41 behavioral measures to predict searchers' perceptions of post-task difficulty [1].

An alternative to studying behavioral measures individually is to *uncover and study the latent phenomena* captured by different measures. In this paper, we report on a secondary analysis of behavioral data gathered from three previously published studies [1, 2, 6]. In each secondary analysis, we used principal component analysis (PCA), a common dimensionality reduction technique, to uncover the latent phenomena captured by a wide range of individual behavioral measures. Additionally, we examined the influence of these latent phenomena on participants' post-task perceptions of workload, difficulty, time pressure, engagement, and knowledge gains. This paper is partly a *tutorial* on how to use PCA to better understand search behaviors and their effects. We demonstrate that applying PCA to behavioral data can provide three important benefits.

First and foremost, PCA can help us understand the underlying phenomena being captured by different behavioral measures. For instance, one underlying phenomenon may relate to *query abandonment*—the extent to which the user/participant issued queries that did not yield good search results. Another underlying phenomenon may relate to *click abandonment*—the extent to which the user/participant clicked on search results that turned out to be not relevant/useful. Finally, a third phenomenon may relate to the searcher's *pace of interaction.*

Secondly, PCA can help us understand a behavioral measure's level of ambiguity with respect to the underlying phenomena being

ICTIR '19, October 2–5, 2019, Santa Clara, CA, USA

ACM ISBN 978-1-4503-6881-0/19/10…$15.00
https://doi.org/10.1145/3341981.3344222

captured. For instance, a measure like "number of queries without clicks" may be strongly related to query abandonment and no other phenomena (i.e., is *unambiguous*). In contrast, a measure like "number of queries" may suggest either query abandonment or general effort. In other words, searchers may issue many queries either because: (1) they are not finding relevant results or (2) the task is complex and demands more information. We show that PCA can help us understand what a behavioral measure "means", and whether it mostly relates to one phenomenon or multiple.

Finally, PCA can help us combine a wide range of behavioral measures into a smaller number of (more coherent) components for further analysis. By doing so, we can understand the effects of the underlying phenomena (vs. individual behavioral measures) on dependent variables of interest (e.g., post-task perceptions).

2 RELATED WORK

A large body of prior work has studied how search behaviors vary based on task characteristics (e.g., complexity), post-task perceptions (e.g., difficulty), and search outcomes (e.g., knowledge gains).

Search Behaviors & Task Characteristics: Athukorala et al. [3] compared participants' behaviors during different task types: fact-finding, navigational, knowledge acquisition, planning, and comparative tasks. The first two task types were considered "look-up" tasks and the last three were consdered "exploratory" tasks. During look-up tasks, participants issued longer queries, performed less scrolling on SERPs, had shorter dwell times on clicked results, and had shorter sessions. Jiang et al. [18] compared participants' search behaviors across tasks that varied based on the task product (factual vs. intellectual) and goal (specific vs. amorphous). The authors found significant differences for a wide range of behavioral measures. For example, during factual (vs. intellectual) tasks, participants had more satisfied SERP clicks (with longer dwell times). Similarly, during amorphous (vs. specific) tasks, participants scanned larger portions of SERPs. Prior work has also considered how search behaviors are affected by a task's *cognitive* complexity—defined as an objective characteristic based on the types (and variety) of mental activities required by the task. Studies have found that cognitively complex tasks require more search activity (e.g., clicks) and have more evidence of trial-and-error (e.g., abandoned queries) [5, 17, 20].

Search Behaviors & Post-task Perceptions: Prior work has focused on determining how search behaviors correlate with different types of post-task perceptions—negative perceptions (e.g., difficulty) and positive perceptions (e.g., engagement).

In terms of negative perceptions, Aula et al [4] compared the behavioral measures of participants who completed tasks successfully vs. unsuccessfully. Unsuccessful participants issued more queries, longer queries, more queries with question words, and spent more time on SERPs. Feild et al. [13] focused on predicting searcher frustration. The authors combined measures related to participants' search interactions with measures derived from a mental state camera and a pressure-sensitive mouse and chair. Interestingly, the interaction measures were more predictive than those derived from the physical sensor data. Liu et al. [21] focused on predicting participants' perceptions of post-task difficulty. Participants who perceived the task to be difficult spent less time between queries, had fewer clicks per query, and spent more time on SERPs.

In terms of positive perceptions, Zhuang et al. [27] used behavioral measures to predict different factors of user engagement. Different behavioral measures were predictive for different factors. For example, participants reported greater *perceived usability* when they spent less time on SERPs and more time reading pages. Similarly, participants reported greater *involvement* when they had more satisfied clicks (with longer dwell times). Edwards and Kelly [11] compared the search behaviors of study participants who were engaged vs. not engaged, as well as participants who were frustrated vs. not frustrated. Engagement was induced by allowing participants to work on tasks they were interested in and frustration was induced by providing poor search results. Interestingly, some measures were found to be *ambiguous*. For example, both highly engaged and frustrated participants performed more SERP scrolls than their un-engaged/un-frustrated counterparts (though likely for different reasons). These results suggest that some behavioral measures may relate to *multiple* phenomena.

Search Behaviors & Knowledge Gains: Collins-Thompson et al. [7] examined the correlations between behavioral measures and participants' knowledge gains during the session. Knowledge gains were positively correlated with the amount of time participants spent viewing non-SERP pages and the number of pages marked useful. In a log-based study, Eickoff et al. [12] compared the search behaviors of users with (procedural or declarative) "knowledge acquisition intent". Results found differences in behaviors during sessions with knowledge acquisition intent versus other sessions. For example, query complexity increased more during knowledge acquisition sessions versus other sessions. Gadiraju et al. [14] compared the search behaviors of study participants with and without knowledge gains, measured using pre-/post-task tests. Several measures had significant positive correlations with knowledge gains: time spent on clicked results, query complexity, and the use of query terms not appearing in the task description.

3 OVERVIEW OF STUDIES

Next, we describe the three studies that generated the data used in our analyses. We refer the reader to the original papers for more details [1, 2, 6]. Studies 1-3 had similarities and differences. In all three studies, participants were assigned search tasks and asked to search for information using custom-built search systems. All systems used search APIs to produce results. Participants' interactions were logged and used to compute the behavioral measures described in Section 4. In all three studies, as the main *goal* of the task, participants were asked to search for and *bookmark* pages that would help them address the task. Finally, participants in all three studies completed post-task questionnaires that measured different perceptions. In terms of differences, Studies 1-3 had different objectives and manipulations. In Study 1, participants completed search tasks of the same type, but interacted with two different interfaces. In Studies 2 & 3, participants used the same system, but completed different *types* of search tasks. In Study 2 we manipulated *task scope* and in Study 3 we manipulated *task complexity*.

3.1 Study 1

Study Overview: Study 1 [2] was a lab-based study with 32 participants. Each participant completed four tasks of the same type and interacted with two different *aggregated search* interfaces that combined search results from various back-end services (or *verticals*). Both aggregated search interfaces combined search results from a web search engine and four vertical search engines: images, news, shopping, and video. Given the same query, both interfaces displayed the same results, but did so in different ways. The *blocked* interface organized results by source: web, images, news, shopping, and video. Specifically, the blocked interface: (1) displayed results from the same source as a group, (2) displayed each group in fixed positions on the SERP regardless of the query, and (3) used visual cues to visually separate results from different sources. In contrast, the *interleaved* interface embedded individual vertical results between the web results in a more "cluttered" fashion. Participants in Study 1 completed four *comparative* tasks, which required comparing/contrasting three specific items (or alternatives) along three specific dimensions (or attributes). Participants were instructed to search for and bookmark *at least* 10 relevant pages in 15 minutes.

Outcome Measures: After each search task, Study 1 participants completed a questionnaire to measure workload. This questionnaire included six items from the NASA-TLX questionnaire [15] to measure: (1) mental demand, (2) physical demand, (3) temporal demand, (4) effort, (5) failure, and (6) frustration. Based on Cronbach's alpha, the internal consistency between these six measures was fairly high ($\alpha = .832$). Therefore, we averaged responses to form one measure of workload. Participants responded using a 7-point scale, with high values indicating high levels of workload.

3.2 Study 2

Study Overview: Study 2 [6] was a crowdsourced study run on Amazon Mechanical Turk (AMT). In this study, 144 participants each completed 6 search tasks, for a total of 864 search sessions. The goal of this study was to investigate the effects of *task scope* on search behaviors and outcomes. Each participant completed 6 comparative tasks (defined in Section 3.1) that varied based on the scope of the task (a within-subjects design). Our manipulation of task scope involved including specific items and/or dimensions for participants to consider as part of the task. Our broadest task version did not specify any items/dimensions for participants to consider, and our narrowest task version asked participants to compare two specific items along one dimension. Additionally, we investigated the effects of asking participants to consider an *objective* vs. *subjective* dimension. We hypothesized that it would be more difficult to address a subjective dimension (e.g., ease of use) than an objective dimension (e.g., cost). Our six task versions were defined as follows: (1) *unspecified*—no items/dimensions specified; (2) *items*—two items specified, no dimension; (3) *objective dimension*—one objective dimension specified, no items; (4) *subjective dimension*—one subjective dimension specified, no items; (5) *items and objective dimension*—two items and one objective dimension specified; and (6) *items and subjective dimension*—two items and one subjective dimension specified. Different from Study 1, Study 2 participants were instructed to bookmark as many pages as necessary (no minimum requirement) and were not imposed any time limits.

Outcome Measures: After each task, Study 2 participants completed a questionnaire to measure engagement. To this end, we used O'Brien's UES-SF questionnaire [25], which measures four engagement factors: (1) focused attention, (2) reward, (3) aesthetic appeal, and (4) perceived usability. The UES-SF is a 12-item questionnaire with 3 items per factor. Using Cronbach's alpha, we measured the internal consistency of each 3-item factor and found acceptable values: focused attention = .846, reward = .886, aesthetic appeal = .923, and perceived usability = .760. Therefore, we averaged responses to form four factors of engagement. Participants responded using a 5-point scale, with high values indicating high engagement.

3.3 Study 3

Study Overview: Like Study 2, Study 3 [1] was a crowdsourced study run on AMT. In this study, participants completed a total of 600 tasks. Participants were randomly assigned tasks of five different levels of *cognitive complexity*, which relates to the types (and variety) of mental activities required to complete the task [5, 17, 20]. In this study, we used the 20 tasks made available by Kelly et al. [20]. These tasks vary across four task topics and five cognitive complexity categories: (1) *remember*—find/verify a specific piece of information; (2) *understand*—construct meaning through synthesis and summarization; (3) *analyze*—break material into parts and determine how the parts relate to each other; (4) *evaluate*—make judgments or assessments through checking and critiquing; and (5) *create*—put elements together to form a novel solution to a problem. Each of the 20 tasks was completed by 30 different AMT workers, for a total of 600 search sessions. Since AMT workers were not allowed the complete the same task more than once, each participant completed 1-20 tasks. Similar to Study 2, participants were instructed to bookmark as many pages as necessary (no minimum requirement) and were not imposed any time limits.

Outcome Measures: In Study 2, we measured participants' post-task perceptions of difficulty, time pressure, and knowledge increase. Participants responded to five difficulty questions: (1) How difficult was it to search for information? (2) How difficult was it to understand the information found? (3) How difficult was it to determine the usefulness of the information found? (4) How difficult was it to decide when you had enough information for the task? and (5) Overall, how difficult was the task? To measure time pressure and knowledge increase, participants responded to one question each: (1) How much time pressure did you feel while working on this task? and (2) How much did your knowledge of the task increase as you searched? The internal consistency between the five difficulty questions was fairly high (Crombach's $\alpha = .903$). Therefore, we averaged responses to form one measure of difficulty. Participants responded using a 5-point scale, with high values indicating high levels of difficulty, time pressure, and knowledge increase.

4 BEHAVIORAL MEASURES

In Studies 1-3, we logged participants' interactions with the search system(s) and computed the following behavioral measures. Not every measure was available for every study, for reasons stated later. We use (1, 2, 3) to denote which studies included each measure.

(1) **Queries:** # of queries issued in the session (1, 2, 3).
(2) **AvgQueryLength:** average query length (in words) (1, 2, 3).
(3) **QuestionQueries:** # of queries with a question word (e.g., what, when) (1, 2, 3).
(4) **QueriesWOBookmarks:** # of queries without a bookmarked page (1, 2, 3).
(5) **QueriesWOClicks:** # of queries without a SERP click (1, 2, 3).
(6) **QueriesWOScrolls:** # of queries without a scroll event (1, 2, 3).
(7) **QueriesWOMouseovers:** # of queries without a mouseover event (1, 2, 3).
(8) **QuickReforms:** # of queries issued ≤ 30 secs. from the previous query (1 ,2, 3).
(9) **RepeatedIntentQs:** # of queries with the same search intent as a previous query (explained below) (1, 2).
(10) **Clicks:** # of SERP clicks (1, 2, 3).
(11) **Views:** # of search results viewed (explained below) (3).
(12) **Bookmarks:** # of pages bookmarked (1, 2, 3).
(13) **ClicksWOBooks:** # of search results clicked and not bookmarked (1, 2, 3).
(14) **ViewsWOBooks:** # of search results viewed and not bookmarked (3).
(15) **AvgClickRank:** average rank across all search results clicked (2, 3).
(16) **AvgViewRank:** average rank across all search results viewed (3).
(17) **AvgBookRank:** average rank across all search results bookmarked (2, 3).

(18) **Paginations:** # of times the participant clicked to see the next page of search results (1, 2, 3).
(19) **ScrollDistance:** total SERP scroll distance in the session, measured in SERP-heights traversed (1, 2, 3).
(20) **TotalMouseovers:** total mouseover events in the session (1, 2, 3)
(21) **AvgMouseoverRank** average rank across all mouseover events (2, 3)
(22) **MouseoversWOClicks:** # of mouseovers without a click (1, 2, 3)
(23) **Avg1stClickTime:** average time (in secs.) between each query and its first SERP click, if any (1, 2, 3).
(24) **TimeToFirstClick:** time (in secs.) to the first click in the session (1, 2, 3).
(25) **TimeToFirstBook:** time (in secs.) to the first bookmark in the session (1, 2, 3).
(26) **AvgDwellTime:** average length (in secs.) of each page view (3).
(27) **TotalDwellTime:** total time (in secs.) spent viewing pages (3).
(28) **TimeBWEvents:** avg. time (in secs.) between subsequent events: queries, clicks, and bookmarks (1, 2, 3).
(29) **CompletionTime:** session length (in secs.) (1, 2, 3).
(30) **UniqueQueries:** # of queries not issued by any other participant (1, 2, 3).
(31) **UniqueQueryTerms:** # of query terms not used by any other participant (1, 2, 3).
(32) **UniqueURLS:** # of SERP clicks not clicked by any other participant (1, 2, 3).

A few of the above measures require more explanation. In Studies 1 and 2, we measured the number of queries with the same search intent as a previously issued query (#9 above). This measure was considered evidence of unsuccessful queries that required reformulation, and was computed by manually coding all queries issued by participants. In Study 3, we logged both SERP clicks and "views". A view is defined as the action of *examining* a clicked result, possibly opened in a new browser tab. For Study 3, we generated several measures based on search results viewed (#11, #14, #16, #26, #27). In Study 1, the blocked and interleaved interfaces had different layouts: the blocked interface used a 2-D grouped layout and the interleaved interface used a 1-D stacked layout. For this reason, in our Study 1 analysis, we did not consider rank-based behavioral measures (#15-#17, #21). Finally, the last three measures (#30-#32) captured the extent to which participants adopted search strategies different from *other* participants who completed the same task.

5 PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA is a dimensionality reduction technique that reduces n variables to m principal components, where $m < n$. In our case, the n variables represent behavioral measures (Section 4). To understand PCA and its applications, it helps to separate the *what* (i.e., "What does PCA do and what does it tell us?") from the *how* ("How do PCA algorithms work?"). In this section, we focus on the *what.*

PCA Intuition: PCA can be viewed as an iterative procedure. The first step is to find the *first* principal component—an n-dimensional (unit) vector representing the direction along which the data has the highest variance. Then, this first principal component is "removed". Effectively, this means that the second principal component (i.e., the direction of the second highest variance) is orthogonal to the first. As the procedure continues, the kth principal component is a variance-maximizing direction *orthogonal* to the previous $k - 1$ components. Dimensionality reduction is achieved by representing the data using the $m < n$ principal components accounting for the most variance (selecting m is discussed later). Ideally, the m principal components will represent *coherent* and *interpretable* phenomena being captured by the original n variables (e.g., behavioral measures). One output from a PCA is an $n \times m$ matrix of component loadings—the correlations between each measure/component pair (e.g., Figures 1-3). Ideally, these loadings provide a means for *interpreting* the m components—by considering which measures have high positive (or negative) loadings with each component. Next, we describe choices that need to be made when performing a PCA. In this paper, our aim was *not* to perform an exhaustive exploration of these choices, but rather to make sensible ones.

Correlation vs. Covariance Matrix: A common approach for performing a PCA is through *eigen decomposition.* In this case, the decomposition is performed on an $n \times n$ matrix describing the relation between all pairs of n variables (in our case, behavioral measures). Here, there are two choices statistical packages may allow: using the covariance or correlation matrix as input. Behavioral measures are often on different scales. For example, the number of queries may be in the tens while the session length (in seconds) may be in the hundreds or thousands. For this reason, we recommend using the correlation (vs. covariance) matrix. The correlation between two variables is in the range [-1,+1] (regardless of scale differences) and is equivalent to their covariance after standardizing each variable to mean of zero and standard deviation of one.

Choosing the Number of Components: PCA requires choosing the number of components (m). Jacobsen and Pedersen [16] recommend the following criteria. First, m should be small compared to n, otherwise nothing is gained from PCA. Secondly, as mentioned above, PCA is based on an eigen decomposition of the input $n \times n$ correlation (or covariance) matrix. The kth eigenvector denotes the direction of the kth principal component, while the kth eigenvalue is the variance of the data when projected onto this component. A common heuristic is to set m equal to the number of eigenvalues > 1. Another is to set m equal to the number of eigenvalues greater than the mean eigenvalue [19]. Finally, a component should not be synonymous with a single variable. That is, every component should have at least two variables with loadings ≥ 0.50 or ≤ -0.50. In this paper, we set m based on the number of eigenvalues greater than the mean eigenvalue.

Rotation: In PCA, rotation is done to better understand component loadings (i.e., how measures relate to components). Oblique rotation techniques allow components to be correlated, while orthogonal rotation techniques favor uncorrelated components. Varimax rotation is a type of orthogonal rotation that maximizes the variance of the squared loadings onto each component. Therefore, it favors loadings that are either very high or very low, making it easier to interpret the underlying phenomenon being captured by each component. In our analyses, we used varimax rotation.

Combining Measures: In order to use PCA to predict outcomes (e.g., task difficulty), the final step is to describe the data using the principal components rather than the original measures. This process requires making two decisions. First, we must decide how to normalize behavioral measures (e.g., standardization, min-max scaling, etc.). A common choice (performed in our analyses) is to standardize measures to mean of zero and standard deviation of one. This is especially appropriate if PCA was performed through eigen decomposition of the $n \times n$ *correlation* (vs. covariance) matrix.

Secondly, we must decide how to combine measures to form components. DiStephano et al. [10] present an overview of different alternatives. In our analyses, we used the *regression* approach [10]. While component *loadings* describe how to linearly combine components to describe each measure, component *scores* describe how to linearly combine measures to form each component (closer to our goal). For each PCA component, the regression approach uses linear regression to estimate each datapoint's projection along the component as a function of the original n measures. Ultimately, the estimated regression coefficients for a given component are called the component *scores* (one per measure), and provide a means to *linearly* combine the n measures to form the component.

6 STATISTICAL ANALYSIS

In Sections 7-8, we present our analysis for Studies 1-3. In Section 7, we first use PCA to group behavioral measures into a few (more coherent) components. Then, in Section 8, we analyze the effects of PCA components on post-task perceptions. To perform this second step, we used multi-level modeling (MLM). MLM is similar to other regression analysis techniques. However, MLM is particularly powerful when the data is "grouped" and we wish to account for *random* effects introduced by different groups. In our case, our data were grouped by participant—Study 1 participants completed 4 tasks each, Study 2 participants completed 6 tasks each, and Study 3 participants completed up to 20 tasks each. Using MLM allowed us to account for random effects at the participant level (e.g., search experience). Specifically, we used random-intercept MLMs. Each MLM (one per outcome variable) had the following equation:

$$Y_{ij} = \underbrace{\beta_0 + \mu_j}_{\textbf{random intercept}} + \underbrace{\beta_1 X_{1i} + \cdots + \beta_m X_{mi}}_{\textbf{fixed factors}} + \underbrace{e_{ijk}}_{\textbf{random error}},$$

where Y_{ij} denotes the outcome variable for datapoint i (associated with participant j). As shown, the y-intercept in our models was a linear combination of β_0 (a global parameter) and μ_j (specific to participant j). Parameters $\beta_1 \ldots \beta_m$ denote the β-values associated with our m components (i.e., the MLM's fixed factors).

To test the significance of each MLM, we computed the χ^2 statistic using the likelihood ratio test against a null model (one without the PCA components as covariates). If a MLM was significant, we performed z-tests to test the significance of each β value. At this, point, we performed Bonferroni correction to account for multiple independent variables (i.e., PCA components). Significantly [positive|negative] β-values indicate a [positive|negative] relation between a component and the outcome variable.

7 RESULTS: PCA OUTPUT

Study 1 PCA Results: For behavioral measures related to Study 1, a PCA with varimax rotation found a six-component solution that explained 76% of the total variance. Table 1 shows the component loadings (i.e., correlation values in the range [-1,+1]) between each behavioral measure and component. In Table 1 (and Tables 2-3), behavioral measures are ordered vertically to emphasize the groupings of measures by component. Additionally, the gray cells indicate loadings that are highly positive (≥ 0.50) or negative (< -0.50). These thresholds are somewhat arbitrary, but are intended to show that some measures loaded strongly with one component while others loaded strongly with multiple (i.e., are more ambiguous). Components are given labels based on our interpretation of the phenomenon being captured. Based on Table 1, our interpretation of the six components for Study 1 is as follows.

- **AbandQs:** PC1 relates to query abandonment—the extent to which a participant issued unsuccessful queries. Measures with high loadings with AbandQs include: (1) number of queries without bookmarks, clicks, mouseovers, and scrolls; (2) number of queries; (3) number of quick query reformulations; (4) number of queries with the same intent as a previous one; and (5) number of queries/query-terms not used by another participant (evidence of an unusual/ineffective query).

Table 1: Study 1 PCA: Component Loadings.

	PC1 (AbandQs)	PC2 (AbandCs)	PC3 (DeepSERP)	PC4 (Pace)	PC5 (NLQs)	PC6 (SlowCs)
QueriesWOBooks	0.87	0.09	0.27	0.04	0.06	-0.03
QueriesWOClicks	0.86	-0.03	0.21	0.21	-0.04	0.02
Queries	0.85	0.16	0.10	-0.29	0.15	0.04
RepeatedIntentQs	0.85	0.14	0.24	-0.02	0.02	0.12
QuickReforms	0.83	0.00	0.09	-0.18	-0.20	-0.05
UniqueQueries	0.82	0.15	0.09	-0.16	0.37	0.00
QueriesWOMouse	0.71	0.02	-0.08	0.15	-0.16	-0.09
UniqueQueryTerms	0.62	0.04	-0.05	0.04	0.36	0.30
QueriesWOScrolls	0.60	0.16	-0.44	-0.33	-0.03	-0.04
Clicks	0.16	0.83	0.22	-0.37	0.07	-0.18
ClicksWOBooks	0.18	0.79	0.25	0.13	0.02	-0.24
UniqueURLs	0.04	0.76	0.08	-0.05	0.12	0.09
CompletionTime	0.05	0.57	0.08	0.53	0.08	0.34
MouseWOClicks	0.34	0.18	0.85	-0.11	0.14	0.06
Mouseovers	0.16	0.40	0.82	-0.09	0.07	-0.02
Paginations	-0.09	0.18	0.78	0.01	-0.09	0.10
ScrollDistance	0.36	-0.06	0.77	-0.03	0.22	0.12
AvgTimeBWEvents	-0.26	-0.05	-0.07	0.83	-0.03	0.31
Bookmarks	-0.02	0.27	-0.07	-0.78	-0.01	0.09
TimeToFirstBook	0.06	0.19	-0.17	0.70	-0.12	0.22
QuestionQueries	0.02	0.08	0.13	0.01	0.83	-0.08
AvgQueryLength	0.01	0.10	0.03	-0.09	0.82	0.10
TimeToFirstClick	0.04	-0.02	0.02	0.12	0.06	0.82
Avg1stClickTime	0.05	-0.13	0.34	0.30	-0.04	0.70

- **AbandCs:** PC2 relates to click abandonment—the extent to which a participant produced unsuccessful clicks. Measures with high loadings with AbandCs include: (1) number of clicks without a bookmark; (2) number of clicks; (3) number of clicked URLs not clicked by another participant (evidence of unusual/ineffective clicks); and (4) task completion time. One interpretation is that more abandoned clicks led to longer completion times.
- **DeepSERP:** PC3 relates to a participant's level of SERP exploration, as evidenced by: (1) mouseover events (with/without clicks); (2) scrolls; and (3) clicks to see the next page of results.
- **Pace:** PC4 relates to a participant's pace of interaction, as evidenced by: (1) average time between subsequent events (i.e., queries, clicks, and bookmarks), (2) time to the first bookmark in the session, and (3) the task completion time. Interestingly, the number of bookmarks had a strong *negative* loading with Pace. This result is possibly an artifact of the experimental design in Study 1—participants were instructed to produce at least 10 bookmarks in 15 minutes. One interpretation is that slower searchers bookmarked fewer pages in the allotted 15 minutes.
- **NLQs:** PC5 relates to the extent to which a participant issued natural language queries (NLQs)—queries with question words and longer queries.
- **SlowCs:** PC6 relates to how long it took a participant to identify potentially relevant results on SERPs. Measures with high loadings with SlowCs include: (1) time to first click in the session and (2) average time between each query and its first click.

The results in Table 1 suggest that some behavioral measures are more ambiguous than others. For example, the number of queries without clicks had a high loading with only one component (AbandQs), while the task completion time had high loadings with multiple (AbandCs and Pace). In other words, a longer completion time may be evidence of more unsuccessful clicks (AbandCs) and/or a slower pace of interaction (Pace). Our PCA results for Studies 2 & 3 also show that some behavioral measures are ambiguous and therefore have strong loadings with multiple components.

Study 2 PCA Results: For behavioral measures related to Study 2, a PCA with varimax rotation found a five-component solution that explained 70% of the total variance. Table 2 shows the component loadings between each behavioral measure and component. Our interpretation of these five components is as follows.

- **AbandQs:** Similar measures as in Study 1.
- **Effort:** PC2 relates to the amount of search effort exerted by a participant. Measures with high loadings with Effort include: (1) number of queries, clicks, and bookmarks; (2) number of clicks without a bookmark; (3) number of mouseovers and scrolls; (4) number of queries not issued by another participant; and (5) number of clicked URLs not clicked by another participant.
- **DeepSERP:** PC3 relates to the level of deep SERP exploration in the form of mouseovers, clicks, and bookmarks at lower ranks, as well as clicks to see the next page of search results.
- **Pace:** PC4 relates to a participant's pace of interaction. Different from Study 1, the number of bookmarks did *not* have a strong loading with Pace. Study 2 participants were not asked to bookmark a minimum number of pages and were not imposed a time limit. Thus, one explanation is that Study 2 participants who interacted at a slower pace did not necessarily bookmark fewer pages. A similar trend was observed in Study 3.
- **NLQs:** Same measures as in Study 1.

Study 3 PCA Results: For behavioral measures related to Study 3, a PCA with varimax rotation found a seven-component solution that explained 76% of the total variance. Table 3 shows the component loadings between each behavioral measure and component. Our interpretation of these seven components is as follows.

- **AbandQs:** Similar measures as in Studies 1 & 2.
- **AbandCs:** PC2 relates to click abandonment—the extent to which a participant produced unsuccessful clicks. Measures with high loadings with AbandCs include: (1) number of clicks/views that did not yield a bookmark and (2) number of clicks/views.
- **DeepSERP:** PC3 relates to the level of deep SERP exploration in the form of mouseovers, clicks, views, and bookmarks at lower ranks, as well as clicks to see the next page of search results.
- **Pace:** PC4 relates to a participant's pace of interaction. Measures with high loadings with Pace include: (1) average time between events; (2) average/total dwell times on viewed results; (3) time to first bookmark; and (4) task completion time.
- **Effort:** PC5 relates to the level of effort exerted by a participant. Measures with high loadings with Effort include: (1) number of views; (2) number of bookmarks; and (3) number of queries and clicks not issued/clicked by another participant.
- **NLQs:** Same measures as in Studies 1 & 2.
- **SlowCs :** PC7 relates to how long it took a participant to identify potentially relevant results on SERPs. Measures with high loadings with SlowCs include: (1) time to first click in the session; (2) average time between each query and its first click; and (3) the total scroll distance. One interpretation is that participants who took longer to click on search results had more scrolls.

8 RESULTS: EFFECTS ON PERCEPTIONS

Next, we report on the effects of PCA components on post-task perceptions. In Table 4, we present results across studies to enable comparisons. In the last column, we report the significance of each MLM using all components from the corresponding study. Again, we computed the significance of each MLM by performing a likelihood ratio (LR) test against a null model. In columns 2-8, we show components that had significant effects (i.e., significant β-values after Bonferroni correction) on the dependent variable (row).

Table 2: Study 2 PCA: Component Loadings.

	PC1	PC2	PC3	PC4	PC5
	(AbandQs)	(Effort)	(DeepSERP)	(Pace)	(NLQs)
QueriesWOBooks	0.94	0.10	0.07	0.08	0.03
QuickReforms	0.88	0.14	-0.01	-0.01	-0.03
QueriesWOBooks	0.86	0.38	0.06	0.13	0.05
Queries	0.82	0.52	-0.02	0.09	0.06
QueriesWOMouse	0.81	-0.11	0.13	0.04	-0.04
RepeatedIntentQs	0.79	0.46	0.04	0.11	0.06
QueriesWOScrolls	0.74	0.01	-0.14	-0.02	0.05
UniqueQueries	0.70	0.59	-0.06	0.05	0.04
Clicks	0.23	0.83	0.31	0.10	0.00
UniqueURLs	0.15	0.79	0.23	0.03	0.10
ClicksWOBooks	0.24	0.74	0.36	0.10	0.00
Mouseovers	0.28	0.71	0.34	0.13	-0.04
MouseWOClicks	0.39	0.68	0.43	0.10	0.04
ScrollDistance	0.22	0.68	0.45	0.17	0.06
Bookmarks	-0.02	0.55	-0.01	-0.03	-0.02
Paginations	0.13	0.53	0.64	0.08	0.08
AvgMouseRank	-0.06	0.26	0.86	0.03	0.00
AvgClickRank	-0.05	0.23	0.86	0.02	-0.05
AvgBookRank	-0.06	0.18	0.81	-0.01	-0.05
AvgTimeBWEvents	-0.10	-0.04	-0.02	0.89	0.06
CompletionTime	0.15	0.40	0.09	0.85	0.03
TimeToFirstBook	0.20	-0.01	0.08	0.82	-0.01
QuestionQueries	0.14	0.02	0.11	-0.04	0.84
AvgQueryLength	-0.13	-0.20	0.16	-0.01	0.78
Avg1stClickTime	-0.05	0.13	-0.16	0.20	0.26
TimeToFirstClick	0.01	0.01	-0.03	0.01	0.10
UniqueQueryTerms	0.44	0.35	-0.11	-0.06	0.36

Table 3: Study 3 PCA: Component Loadings.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
	(AbandQs)	(AbandCs)	(DeepSERP)	(Pace)	(Effort)	(NLQs)	(SlowCs)
QueriesWOClicks	0.91	-0.05	0.12	0.01	0.03	0.06	0.08
QueriesWOBooks	0.90	0.27	0.03	0.05	0.05	0.00	0.03
QuickReforms	0.83	0.11	0.30	-0.02	0.07	-0.07	0.08
Queries	0.82	0.28	0.01	0.03	0.41	-0.02	0.01
QueriesWOScrolls	0.79	0.09	-0.13	-0.07	0.27	0.02	0.02
QueriesWOMouse	0.78	-0.02	0.05	0.03	-0.09	0.05	0.15
ClicksWOBooks	0.22	0.87	0.17	0.12	-0.01	-0.05	-0.08
ViewsWOBooks	0.10	0.87	0.14	0.09	-0.02	-0.04	-0.04
Clicks	0.18	0.75	0.33	0.06	0.41	-0.06	-0.10
Views	0.09	0.69	0.34	0.02	0.50	-0.07	-0.08
AvgMouseRank	0.05	0.12	0.94	0.06	0.09	-0.04	0.11
AvgViewRank	-0.03	0.08	0.93	0.05	0.03	0.09	-0.06
AvgClickRank	0.00	0.07	0.93	0.06	0.04	0.07	0.09
AvgBookRank	-0.03	0.10	0.92	0.05	0.02	0.10	-0.05
Paginations	0.13	0.14	0.76	0.00	0.06	-0.15	0.19
Mouseovers	0.18	0.40	0.71	0.05	0.32	-0.13	0.03
MouseWOClicks	0.33	0.26	0.63	0.03	0.32	-0.12	0.06
AvgTimeBWEvents	-0.05	-0.14	0.02	0.89	-0.02	-0.03	0.13
AvgDwellTime	-0.06	-0.10	-0.04	0.88	-0.05	-0.01	-0.04
TotalDwellTime	0.00	0.32	0.18	0.74	0.25	-0.04	-0.07
TimeToFirstBook	0.15	0.31	0.04	0.74	-0.13	0.03	0.25
CompletionTime	0.17	0.42	0.31	0.60	0.45	-0.03	0.08
Bookmarks	0.03	0.17	0.37	-0.05	0.78	-0.06	-0.08
UniqueQueries	0.33	0.10	-0.07	0.06	0.65	-0.04	0.08
UniqueURLs	0.09	0.47	0.38	0.06	0.50	0.08	-0.01
QuestionQueries	0.18	-0.07	-0.02	-0.02	0.11	0.81	-0.04
AvgQueryLength	-0.11	0.06	0.00	-0.03	-0.19	0.80	0.04
TimeToFirstClick	0.28	-0.10	0.08	0.09	0.00	-0.01	0.85
Avg1stClickTime	-0.05	0.04	0.06	0.19	0.03	0.09	0.76
ScrollDistance	0.10	0.00	0.06	-0.04	-0.03	-0.08	0.71
UniqueQueryTerms	-0.01	0.29	0.01	-0.03	0.11	0.12	0.14

Workload (Study 1): In Study 1, all six components found by PCA had significant *positive* effects on perceived workload. Greater levels of workload were reported when participants had more unsuccessful queries (AbandQs); more unsuccessful SERP clicks (AbandCs); more SERP-level exploration (DeepSERP); had a slower pace of interaction (and therefore fewer bookmarks) (Pace); issued more natural language queries (NLQs); and took longer to click on search results (SlowCs).

Table 4: Effects of PCA components on dependent variables related to Study 1-3. Symbols 'ns' denotes 'not significant', '*' denotes $p < .001$, '**' denotes $p < .01$, and '*' denotes $p < .05$. Symbol '–' denotes that a component was not available for a specific study. The values in columns AbandQ—Effort correspond to significant β-values in the corresponding MLM (row).**

Study/dependent variable	**AbandQ**	**AbandC**	**DeepSERP**	**Pace**	**NLQs**	**SlowC**	**Effort**	**LR test** vs. null
Study 1								
workload	$.32^{***}$	$.27^{**}$	$.26^{**}$	$.52^{***}$	$.19^{*}$	$.24^{**}$	–	$\chi^2(6)=72.17^{***}$
Study 2								
focused attention	*ns*	–	*ns*	*ns*	*ns*	–	*ns*	*ns*
reward	$-.08^{**}$	–	*ns*	*ns*	*ns*	–	$-.11^{***}$	$\chi^2(5)=28.89^{***}$
aesthetic appeal	$-.05^{**}$	–	*ns*	*ns*	*ns*	–	*ns*	$\chi^2(5)=18.75^{**}$
perceived usability	$-.18^{***}$	–	$-.12^{***}$	*ns*	*ns*	–	$-.21^{***}$	$\chi^2(5)=119.96^{***}$
Study 3								
difficulty	$.13^{***}$	$.23^{***}$	$.15^{***}$	*ns*	*ns*	*ns*	$.10^{*}$	$\chi^2(7)=90.94^{***}$
time pressure	$.10^{**}$	$.23^{***}$	*ns*	$.13^{**}$	*ns*	*ns*	$.14^{***}$	$\chi^2(7)=86.47^{***}$
knowledge increase	*ns*	*ns*	*ns*	.13 ($p = .052$)	*ns*	*ns*	*ns*	$\chi^2(7)=21.60^{***}$

Engagement (Study 2): In Study 2, three of the five components found by PCA had significant *negative* effects on engagement measures. Lower reward was reported when participants had more unsuccessful queries (AbandQs) and exerted more effort (Effort). Lower aesthetic appeal was reported when participants had more unsuccessful queries (AbandQs). Finally, lower perceived usability was reported when participants had more unsuccessful queries (AbandQs), deeper SERP-level exploration (DeepSERP), and exerted more effort (Effort). No components predicted focused attention.

Difficulty, Time Pressure, Knowledge Increase (Study 3): In Study 3, five of seven PCA components had an effect on the dependent measures. Participants perceived the task to be more difficult when they had more unsuccessful queries (AbandQs), more unsuccessful SERP clicks (AbandCs), deeper SERP-level exploration (DeepSERP), and exerted more effort (Effort). Participants reported more time pressure when they had more unsuccessful queries (AbandQs), more unsuccessful SERP clicks (AbandCs), had a slower pace of interaction (Pace), and exerted more effort (Effort). Finally, Pace had a marginally significant effect on knowledge increase ($p = .052$). This result suggest that participants reported greater knowledge gains when they took longer between search events, had longer page dwell times, and took longer to complete tasks.

9 DISCUSSION & CONCLUSION

Next, we summarize our results and discuss their implications.

Untangling Behavioral Measures: Our results show that PCA is a useful tool for understanding the latent phenomena captured by behavioral measures. In this respect, our results suggest four important advantages of a PCA approach. First, our results suggests that search sessions are characterized by similar phenomena. Four components (AbandQs, DeepSERP, Pace, NLQ) were common to all three studies, and every component was common to at least two studies. While this may not be surprising, it suggests that future studies should consider behavioral measures that capture these phenomena—query abandonment, click abandonment, deep SERP exploration, general search effort, natural language queries, interaction pace, and speed in finding relevant results after a query.

Secondly, our results suggest that PCA can help us distinguish between behavioral measures that are unambiguous versus ambiguous with respect to latent behavioral phenomena. Based on our results, the "number of queries without clicks" is an example of an *unambiguous* measure. Across Studies 1-3, this measure had the strongest loading with AbandQs (extent of unsuccessful querying) and weak loadings with all other components. Conversely, across Studies 1-3, we found several examples of *ambiguous* measures. In Study 1, "task completion time" had strong loadings with AbandCs and Pace, suggesting that participants took longer to complete tasks when they had more abandoned clicks and/or simply interacted at a slower pace. In Study 2, "number of queries" had strong loadings with AbandQs and Effort, suggesting that participants issued more queries when they had more abandoned queries and/or exerted more effort (e.g., the task required more information). Finally, in Study 3, "number of pages viewed" had strong loadings with AbandCs and Effort, suggesting that participants viewed more pages when they had more abandoned clicks and/or exerted more effort (e.g., the task required more information). These results suggest that PCA can help reveal which measures have ambiguous/unambiguous interpretations.

Third, our results suggest that a study's experimental design can influence the types of latent phenomena related to participants' behaviors. To illustrate, different from Study 1, Studies 2 & 3 involved a task manipulation (task scope and complexity, respectively). Importantly, in both original studies, certain task types were perceived to be more difficult and required more search activity [1, 6]. In this paper, one of the PCA components found for Studies 2 & 3 was interpreted as Effort. In both studies, measures related to Effort included the number of results examined and bookmarked. Our interpretation of this result is that certain tasks in Studies 2 & 3 required more information. On the other hand, Study 1 did *not* involve a task manipulation. Possibly for this reason, for Study 1, PCA did not reveal a component analogous to Effort. In other words, Study 1 participants exerted similar amounts of search effort, but their behaviors varied in other ways.

The final advantage of PCA is a continuation of the previous point. A study's experimental design can influence, not only the latent phenomena captured by behavioral data, but also an individual measure's *interpretation.* Consider a measure such as "number of bookmarks". In Study 1, participants were instructed to bookmark at least 10 pages in 15 minutes. Conversely, in Studies 2 & 3, participants were instructed to bookmark *any number* of pages and were *not* imposed a time limit. In Study 1, the number of bookmarks had a strong *negative* loading with Pace, suggesting that slower participants (e.g., who took longer between events) bookmarked fewer pages in the allotted 15 minutes. Conversely, in Studies 2 & 3, the number of bookmarks loaded strongly with Effort instead of Pace. In other words, in Studies 2 & 3, participants bookmarked more pages when they exerted more effort (i.e., the task demanded more information) *regardless* of their pace of interaction. This trend suggests that PCA can help us interpret measures that may "mean" different things depending on the experimental design.

Effects on Post-task Perceptions: PCA can not only help us interpret behavioral data, it can also help us understand the impact of different behavioral phenomena (versus individual measures) on post-task perceptions. Here, we summarize our results and compare them to those from previous studies.

In terms of negative perceptions, *all* Study 1 components had a significant positive effect on workload. Participants reported more workload when they had more query abandonment (AbandQs), click abandonment (AbandCs), deeper SERP exploration (DeepSERP), interacted at a slower pace (Pace), issued longer queries with question words (NLQ), and took longer to click on search results (SlowCs). In terms of difficulty, three Study 3 components had significant positive effects. Participants reported more difficulty when they had more query abandonment (AbandQs), click abandonment (AbandCs), and exerted more effort (Effort). These results are consistent with Liu et al. [21]. In that study, participants reported more difficulty when they had quicker query-reformulations and fewer clicks per query (i.e., evidence of ineffective querying). Finally, in terms of time pressure, four Study 3 components had significant positive effects. Participants reported greater levels of time pressure when they had more query abandonment (AbandQs), click abandonment (AbandCs), interacted at a slower pace (Pace), and exerted more effort (Effort). This result is consistent with Crescenzi et al. [8]. In that study, there was a significant positive correlation between the task completion time and participants' perceptions of time pressure.

In terms of positive perceptions, three Study 2 components had significant negative effects on factors of engagement. Participants reported lower engagement when they had more query abandonment (AbandQs), deeper SERP exploration (DeepSERP), and exerted more effort (Effort). This result is consistent with Zhuang et al. [27]. In that study, participants reported lower engagement when they spent more time searching and less time reading pages. Finally, in terms of knowledge gains, one Study 3 component had a *marginally* significant positive effect: Pace. Participants reported greater knowledge gains when they interacted at a slower pace and spent more time reading pages. Similar results were found in Collins-Thompson et al. [7] and Gadiraju et al. [14]. In those studies, there was a positive correlation between participants' knowledge gains and the time spent reading pages (vs. searching).

Concluding Remarks: In this paper, we reported on three secondary analyses of data from three previously published studies. In each secondary analysis, we used principal component analysis (PCA) to study the latent phenomena captured by a wide range of behavioral measures. Additionally, we examined the influence of these latent phenomena (vs. individual measures) on participants' post-task perceptions of workload, difficulty, time pressure, engagement, and knowledge gains. Our results suggest that applying PCA to behavioral data provides several important benefits. First, PCA can help us understand the latent phenomena being captured by different behavioral measures. In this respect, our results suggest that a study's experimental design can influence the latent phenomena uncovered *and* an individual measure's interpretation. Second, PCA can help us understand a behavioral measure's level of ambiguity—the extent to which it relates to one phenomenon or multiple. Finally, by combining behavioral measures into a few (more coherent) components, PCA can help us understand the relation between behavioral phenomena and search outcomes. An important question in IIR research is: What do search behaviors tells about specific user/task characteristics and perceptions of the search experience? In this paper, we have argued that IIR research should consider PCA as a useful tool for interpreting behavioral measures and studying their effects.

Acknowledgements: This work was supported by NSF grant IIS-1451668.

REFERENCES

[1] Jaime Arguello. 2014. Predicting Search Task Difficulty. In *ECIR*. Springer, 88–99.

[2] Jaime Arguello and Bogeum Choi. 2019. The Effects of Working Memory, Perceptual Speed, and Inhibition in Aggregated Search. *TOIS* 37, 3 (2019), 1–34.

[3] K. Athukorala, Dorota G, Giulio Jacucci, Antti Oulasvirta, and Jilles Vreeken. 2016. Is Exploratory Search Different? A Comparison of Information Search Behavior for Exploratory and Lookup Tasks. *JASIST* 67, 11 (2016), 2635–2651.

[4] Anne Aula, Rehan M. Khan, and Zhiwei Guan. 2010. How Does Search Behavior Change As Search Becomes More Difficult?. In *CHI*. ACM, 35–44.

[5] Robert Capra, Jaime Arguello, Anita Crescenzi, and Emily Vardell. 2015. Differences in the Use of Search Assistance for Tasks of Varying Complexity. In *SIGIR*. ACM, 23–32.

[6] Robert Capra, Jaime Arguello, Heather O'Brien, Yuan Li, and Bogeum Choi. 2018. The Effects of Manipulating Task Determinability on Search Behaviors and Outcomes. In *SIGIR*. ACM, 445–454.

[7] Kevyn Collins-Thompson, Soo Young Rieh, Carl C. Haynes, and Rohail Syed. 2016. Assessing Learning Outcomes in Web Search: A Comparison of Tasks and Query Strategies. In *CHIIR*. ACM, 163–172.

[8] Anita Crescenzi, Robert Capra, and Jaime Arguello. 2017. Time Limits, Information Search and the Use of Search Assistance. In *CHIIR*. ACM, 349–352.

[9] Anita Crescenzi, Diane Kelly, and Leif Azzopardi. 2016. Impacts of Time Constraints and System Delays on User Experience. In *CHIIR*. ACM, 141–150.

[10] Christine Distefano, Min Zhu, and Diana Mindrila. 2008. Understanding and Using Factor Scores: Considerations for the Applied Researcher. *Pract. Assess. Res. Eval.* 14 (11 2008).

[11] Ashlee Edwards and Diane Kelly. 2017. Engaged or Frustrated?: Disambiguating Emotional State in Search. In *SIGIR*. ACM, 125–134.

[12] Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. 2014. Lessons from the Journey: A Query Log Analysis of Within-session Learning. In *WSDM*. ACM, 223–232.

[13] Henry A. Feild, James Allan, and Rosie Jones. 2010. Predicting Searcher Frustration. In *SIGIR*. ACM, 34–41.

[14] Ujwal Gadiraju, Ran Yu, Stefan Dietze, and Peter Holtz. 2018. Analyzing Knowledge Gain of Users in Informational Search Sessions on the Web. In *CHIIR*. ACM.

[15] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*. Advances in Psychology, Vol. 52. North-Holland, 139–183.

[16] Judith L. Jacobsen and Jesper Pedersen. 1995. Principal Component Analysis of Behavioral Data: A Case Study of Previously Presented Data. *Nordic Journal of Psychiatry* 49, 6 (1995), 447–457.

[17] Bernard J. Jansen, Danielle Booth, and Brian Smith. 2009. Using the Taxonomy of Cognitive Learning to Model Online Searching. *I&PM* 45, 6 (2009), 643–663.

[18] Jiepu Jiang, Daqing He, and James Allan. 2014. Searching, Browsing, and Clicking in a Search Session: Changes in User Behavior by Task and over Time. In *SIGIR*. ACM, 607–616.

[19] I. T. Jolliffe. 1986. *Principal Component Analysis and Factor Analysis.* Springer New York, New York, NY, 115–128.

[20] 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 *ICTIR*. ACM, 101–110.

[21] Chang Liu, Jingjing Liu, and Nicholas J. Belkin. 2014. Predicting Search Task Difficulty at Different Search Stages. In *CIKM*. ACM, 569–578.

[22] Jingjing Liu, Michael J. Cole, Chang Liu, Ralf Bierig, Jacek Gwizdka, Nicholas J. Belkin, Jun Zhang, and Xiangmin Zhang. 2010. Search Behaviors in Different Task Types. In *JCDL*. ACM, 69–78.

[23] Jingjing Liu, Chang Liu, Jacek Gwizdka, and Nicholas J. Belkin. 2010. Can Search Systems Detect Users' Task Difficulty?: Some Behavioral Signals. In *SIGIR*. ACM, 845–846.

[24] Matthew Mitsui, Jiqun Liu, and Chirag Shah. 2018. The Paradox of Personalization: Does Task Prediction Require Individualized Models?. In *CHIIR*. ACM, 277–280.

[25] Heather L O'Brien, Paul Cairns, and Mark Hall. 2018. A Practical Approach to Measuring User Engagement with the Refined User Engagement Scale (UES) and New UES Short Form. *International Journal of Human-Computer Studies* (2018).

[26] Ran Yu, Ujwal Gadiraju, Peter Holtz, Markus Rokicki, Philipp Kemkes, and Stefan Dietze. 2018. Predicting User Knowledge Gain in Informational Search Sessions. In *SIGIR*. ACM, 75–84.

[27] Mengdie Zhuang, Gianluca Demartini, and Elaine G. Toms. 2017. Understanding Engagement Through Search Behaviour. In *CIKM*. ACM, 1957–1966.