Evaluating Search Explainability with Psychometrics and Crowdsourcing

Catherine Chen
catherine_s_chen@brown.edu
Brown University
Providence, Rhode Island, USA

Carsten Eickhoff
c.eickhoff@acm.org
University of Tübingen
Tübingen, Germany

ABSTRACT

Information retrieval (IR) systems have become an integral part of our everyday lives. As search engines, recommender systems, and conversational agents are employed across various domains from recreational search to clinical decision support, there is an increasing need for transparent and explainable systems to guarantee accountable, fair, and unbiased results. Despite many recent advances towards explainable AI and IR techniques, there is no consensus on what it means for a system to be explainable. Although a growing body of literature suggests that explainability is comprised of multiple subfactors, virtually all existing approaches treat it as a singular notion. In this paper, we examine explainability in Web search systems, leveraging psychometrics and crowdsourcing to identify human-centered factors of explainability.

1 INTRODUCTION

Artificial Intelligence (AI) and Machine Learning (ML) systems have become ubiquitous, in part due to their ability to efficiently synthesize large amounts of data to achieve high performance on complex problems. Corporations are able to detect market trends, maximize profit margins, or deliver personalized content tailored to individual preferences.

Although these integrations have become increasingly popular and show convincing performance, there are cases where AI/ML systems have failed publicly in concerning ways [2, 50]. These failures highlight a serious need for understanding how highly parametric non-linear models are making their decisions so that we can identify where these models might fail and adjust accordingly. As a result, explainable AI[1] initiatives and regulations have emerged in an attempt to instill more trust in AI/ML systems by creating techniques to explain model decisions in human understandable terms and provide more transparency [27, 29].

While explainable search systems have been studied in the IR community for some time, spanning from early approaches such as Tile-Bars [31] to more recent work such as uRank [18] and EXS [57], there is little research on quantifying the degree of explainability attained by these systems. Current approaches toward evaluating explainable ML and IR systems fall short in two ways. Firstly, despite a growing body of literature that suggests the multidimensionality of explainability [21, 43, 48], there is no clear consensus which concrete aspects should contribute to the definition. Secondly, evaluation invariably occurs on a binary scale, in the sense that a system is declared either explainable or a black-box. To address these shortcomings, we (1) identify individual factors of explainability through factor analysis, and (2) combine these factors to propose a multidimensional definition Web search explainability.

Inspired by previous work on multidimensional relevance modeling [69], we leverage *psychometrics* [24] and crowdsourcing to do so. Psychometrics is a well-established field of study in psychology used to develop assessments or measurement models of cognitive constructs which cannot be measured directly. Our psychometric methodology was implemented in several phases. First, we conducted a comprehensive literature review and identified a broad range of explainability aspects that have been well-discussed in the community. Next, we designed a user study to confirm a multidimensional model, as psychometric methodologies are data-driven techniques and crowdsourcing is an efficient way to collect data. Additionally, since Web search does not assume the searcher to have advanced degrees of domain-specific knowledge, crowdsourcing is an appropriate method of gathering diverse results for the everyday layperson. Finally, we used the outcomes from our crowdsourced study to introduce a multidimensional definition of explainability.

The work outlined in this paper aims to create novel techniques to advance the state of knowledge in explainable search systems with the goal of empowering users to understand the processes that cater to their daily information needs in an environment potentially fraught with biases, and misinformation. Specifically, we make the following contributions:

- Leverage psychometrics and crowdsourcing to test well-discussed aspects of explainable Web search systems from the literature that users find most important
- Introduce a hierarchical two-factor definition of Web search explainability using structural equation modeling

The remainder of this paper is structured as follows: In Section 2, we present background and related work on psychometric studies, the multidimensionality of explainability, and previous attempts to evaluate explainable systems. In Section 3, we outline steps taken to develop our measuring instrument and crowdsourcing task setup. In Section 4, we present the results of our data collection and model creation efforts. Finally, in Section 5, we analyze the dimensions of explainability users found important and discuss the limitations of our study. Section 6 concludes with an overview of future work.

2 RELATED WORK

Drawing from previous work to create multidimensional models of relevance [68, 69], we leverage psychometrics and crowdsourcing to establish several factors that contribute to the definition of explainability to create a multidimensional model.

[1]Explainable AI (XAI) and interpretable machine learning are closely related terms, used alongside the notions of explainability and interpretability. In this paper, we follow accepted literature practice and use them to refer to the act of providing some insight into model decision making processes [1, 11, 20, 48].

2.1 Psychometrics, SEM, and Crowdsourcing

Psychometrics is a branch of psychology that examines the theory and techniques of psychological measurement, in particular, the validity and reliability of tests that are constructed to measure cognitive properties [24]. While we cannot directly measure psychological constructs, we can ask questions that represent their manifestations and use Structural Equation Modeling (SEM) to construct a model from the observed relationships in the data.

SEM is a well-established method within psychometrics used to measure the presence of latent and observed variables and analyze the relationships between them [63]. *Latent variables* (factors) are representations of cognitive properties that can be measured through *observed variables* (questionnaire items).

SEM is composed of two parts: (1) *Exploratory Factor Analysis (EFA)* to produce a hypothesized model structure from observed data and (2) *Confirmatory Factor Analysis (CFA)* to confirm the EFA-derived model fit on a held-out dataset. EFA is used to determine the number of latent factors and which items load on the discovered dimensions. It is a statistical technique that first assumes all items load on individual factors, then items are iteratively grouped and pruned until the variable set is reduced to reach a high-quality estimate of covariance in the observed data set. CFA is used to validate the EFA-derived model's fit. Model parameters are re-estimated using maximum likelihood on a held-out set of observed data and model fit is assessed via statistical significance testing. Since significance testing can only reject or fail to reject a model, CFA involves testing multiple alternative models (often including the null model) to determine best fit.

Since SEM requires large amounts of user response data, crowdsourcing is often used for data collection due to its convenience and efficiency in quickly recruiting a large number of participants. Although a variety of crowdsourcing platforms exist, our choice, Amazon Mechanical Turk (MTurk), is the most widely used by researchers. However, one of the biggest challenges in crowdsourcing is collecting quality data, since the payout may be the main motivator for workers to complete tasks and platforms become more saturated with low-quality workers. To mitigate these issues, preventative measures can be taken by setting high worker qualifications, enabling rigorous quality control checks, and post-processing data for inattentive responses to verify quality work [6, 26, 37, 46]. We describe the quality control checks we employ during our study in Section 3 and throughout the course of our study, we received feedback from workers showing promising engagement and understanding of our task.

2.2 Dimensions of Explainability

The term "black box" is often used to refer to ML models whose decision-making processes are not easily understandable to humans. Existing work on explainability aims to make these internal processes more transparent. Yet, explainability is still often considered to be a binary concept. In other words, if a system is not a complete "black box", then it is deemed to be explainable and vice versa. However, recent literature suggests that complex composite concepts such as explainability may be best measured as a combination of several factors, rather than a single notion [22, 43, 48].

Lipton [43] and Doshi-Velez and Kim [21] suggest that the concept of explainability is (1) ill-defined with no consensus and (2) an amalgamation of several factors rather than a monolithic concept. Specifically, both papers recognize the need to ground the explainability in the context of certain desiderata, such as *trustworthiness* or *causality*. Nauta et al. [48] additionally identify twelve such conceptual properties for the systematic evaluation of explainability as a multidimensional concept. In this work, we take a step towards answering their calls for quantifiable evaluation methods and introduce a data-driven approach to discover latent factors of explainability. Although our approach is targeted toward text-based search, our methodology can be applied to other contexts outside of IR.

3 STUDY DESIGN

Our psychometric methodology consists of four phases: (1) questionnaire design, (2) data collection, (3) exploratory factor analysis (EFA), and (4) confirmatory factor analysis (CFA). In this section, we detail our process to develop our questionnaire through an extensive literature review and our data collection task setup.

3.1 Questionnaire Design

The goal of our questionnaire was to determine which aspects are present and contribute to the overall definition of explainability and quantify the degree of contribution. Although cognitive traits can not be measured directly, we can ask questions that represent manifestations of these aspects [24].

First, to compile a list of candidate aspects that may potentially contribute to the composite notion of explainability, we conducted a comprehensive structured literature review through Google Scholar, aiming to include any well-discussed explainability aspects posited by the community. We included the proceedings of ML, IR, natural language processing (NLP), and human-computer interaction (HCI) venues and noted papers for further review if titles included the keywords *interpretability*, *explainability*, or *transparency*, and cross-referenced papers using *connectedpapers.com* to find similar papers, resulting in 44 papers (37 of which were published within the last 7 years). We then read abstracts and conclusions for this pool to retain only those papers that examined some concrete element or aspect of explainability/interpretability, leaving us with 14 papers covering 26 unique aspects of explainability (i.e., *trustworthiness*, *uncertainty*, *faithfulness*) (Table 1). Our final number of candidate aspects is consistent with, and perhaps more encompassing than, other survey papers such as Nauta et al. [48], who find 12 explainability factors from the literature. Given the flexibility of our framework, future work could easily investigate additional aspects from broader literature.

Next, these aspects were turned into a set of concrete questions (referred to as "items" in psychometrics) to be included in the questionnaire. We recorded responses to these items on a seven-point Likert scale ranging from 1 (Strongly Disagree), via 4 (Neutral), to 7 (Strongly Agree). Our questionnaire was created using the following guidelines [23]: (1) items should use clear language and avoid complex words, (2) items should not be leading or presumptuous, and (3) the instrument should include both positively and negatively keyed items.

Table 1: Candidate Aspects

Aspect	Definition	Sources
Simulatability	Ability to step through a system w/o a computer for a given input and produce the correct output	[3, 43, 58]
Decomposability	Each part of a system's components can be understood and explained	[3, 43]
Algorithmic Transparency	Understanding the system's learning algorithm	[3, 43]
Causality	Ability to infer causal relationships from observational data	[3, 43]
Uncertainty	How confident the model is in its prediction	[8, 21, 22]
Immediacy	When a search query is modified, how quickly outcomes of the query are displayed to the user	[56]
Visibility	When a search query is slightly modified, how changes in the ranking are presented to the user	[53, 56]
Transferability	Ability to use system in different search contexts	[3, 43]
Model Fairness	Model makes fair/ethical decisions	[3, 43]
Understandability	Result interface presents rankings in a manner that can be understood easily by users	[9]
Informativeness	Provides useful information for task	[3, 43]
Global Interpretability	Knowing general factors that contribute to ranking results	[21, 22, 53]
Local Interpretability	Knowing the reasons for specific rankings	[21, 22, 53]
Counterfactuals	Ability to correctly determine how small changes to a query will affect ranking results	[58]
Efficiency	Time spent understanding the interface	[9]
Criticism	Knowing where and how the search engine may fail to explain certain data points	[39]
Compositionality	Structure of result interface	[21, 22, 42]
Units of Explanation	Form and number of cognitive chunks	[21, 22, 42]
Acceptability	Accepted for use	[9]
Faithfulness	How accurately the interface reflects the true reasoning process of the search engine	[36]
Plausibility	How convincing the ranking results are to users	[36]
Accuracy	How well the interface describes how the search engine ranked the results	[36]
Completeness	Result interface provides accurate and complete descriptions of the search engine's operations	[25]
Trustworthiness	Confidence in ranking result accuracy	[3, 43, 52]
Justifiability	System produces results that align with human expert judgements	[9]
Explanation Fairness	System is accessible and fair towards all people	[36]

Additionally, as explainability relies on both system and explanation perception, we created items taking both into account, so our final evaluation would reflect these desiderata. To combat fatigue effects, we chose to create 2 items per aspect (one positively and one negatively worded), for a total of 52 items presented in fully randomized order, with the expectation that the discovery of latent factor representations during factor analysis would establish groupings of multiple related items. 11 doctoral and post-doctoral researchers reviewed our questionnaire for clarity and accuracy given aspect definitions. From this evaluation, we were able to identify and correct potential inconsistencies before our pilot study.

3.2 Task Setup

We asked participants to perform a series of 3 search tasks distributed across 3 topics. To motivate and guide their search, users were asked to answer a multiple choice question for each topic and we provided them with a mock search engine that displayed a query and a list of results. The search interface was based on current commercial search engines. We hosted our site on Netlify and displayed the task on MTurk.

Topics and questions were selected from the TREC 2004 Robust Track Dataset [64], which contains a collection of documents from the Federal Register, Financial Times, Foreign Broadcast Information Service, and LA Times. Multiple choice answers were created by the authors such that all answers would not be found on the first page of results and required clicking into documents. To ensure that we would have enough relevant documents to populate the results page, we randomly sampled 9 topics that contained at least 50 relevant documents. Topics were grouped for diversity as follows: **(A)** industrial espionage; income tax evasion; in vitro fertilization, **(B)** radioactive waste; behavioral genetics; drugs in the Golden Triangle, **(C)** law enforcement, dogs; non-US media bias; gasoline tax in US. For each topic, we presented 100 pre-selected documents with a 50/50 random sample of relevant and non-relevant documents in a randomized ranking order so workers would be required to interact with the search system in order to successfully complete the multiple choice quiz.

Participants were randomly assigned a topic grouping and after completing the search task, we presented each group with another mock search interface, and participants were asked to complete our questionnaire for this second search system. The search interfaces

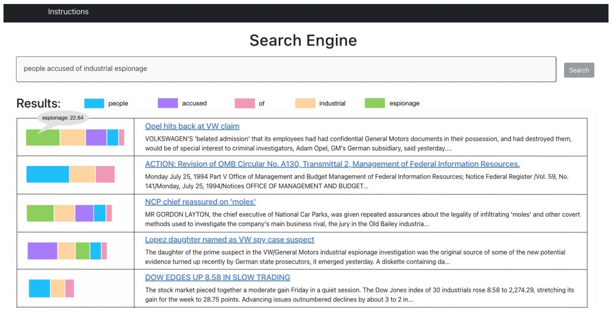

Figure 1: Group A search interface (modeled on the basis of the system presented by Ramos and Eickhoff [54]). On the left-hand side, the stacked bar graphs depict hypothetical scores of each keyword in the query for each respective search result. The larger the stacked bar graph, the more relevant that result is to the query.

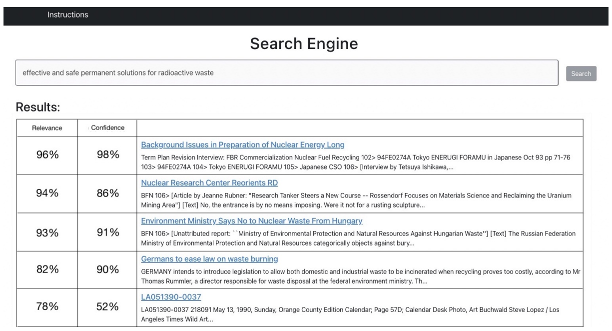

Figure 2: Group B search interface (modeled on the basis of the system presented by Cohen et al. [13]). On the left-hand side, the two columns show a hypothetical relevance and confidence score for each result. The relevance score refers to how relevant the corresponding article is to the query. The confidence score refers to how confident the search engine is that it correctly calculated the relevance score.

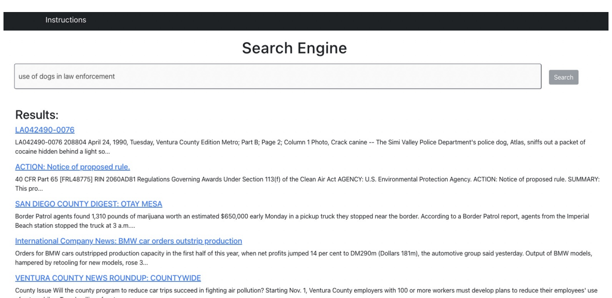

Figure 3: Group C search interface. This interface served as our baseline as an example of a system with limited explainable elements.

were modeled on the basis of existing transparent search systems from the literature [13, 54] to test systems of varying degrees of explainability. To avoid priming effects and other potential biases, we employed a between-subjects study design. Participants in Group **A** were presented with an interface that provided visual explanation aids, with stacked bar graphs displayed next to each search result that informed users how much each query term influenced the corresponding document ranking (Figure 1). Participants in Group **B** were presented with an interface that displayed relevance and confidence scores for each result, where confidence was modeled as a function of uncertainty (Figure 2). Participants in Group **C** were presented with the same non-transparent system they interacted with during the screening search task (Figure 3). Each condition was accompanied by brief usage instructions explaining the novel (if any) interface features.

We included multiple quality control checks to verify worker attentiveness and effort on our task. Specifically, we monitored site interactions (number of clicks, documents viewed, time spent on task), employed a multiple choice quiz, and provided a unique code for the worker to submit on MTurk's site to verify the successful completion of our task. In addition to serving as a form of quality control, the multiple choice quiz was employed to help guide the workers through the search task and put them into a search mindset so they could more accurately complete the questionnaire.

While users' familiarity with topics might impact their experience during the search task, the main results of this study are drawn from the questionnaire experience, which (1) was separate from the search task where the topics were presented and (2) users were asked to comment on the nature of a system, not the search task they previously performed. The questionnaire was intended to capture the extent of perceived system explainability.

3.3 Pilot Study

The goal of our pilot study was to assess the feasibility of our task, determine final worker qualifications, and make any necessary adjustments to the experiment workflow. We collected a total of 62 responses from MTurk workers over a two-month period.

Observations made from user behavior during the pilot study influenced a number of changes in our task design. We found that although we enumerated several specific scenarios that would lead to rejection (i.e., failing the multiple choice quiz, zero interactions with the interface, etc.), some workers would misread or skip the instructions, so we added several pop-up confirmations to remind workers of the experiment rules before they began the task. We also found that requiring a unique completion code served as an effective quality control measure and made it easier to automatically reject workers who did not complete our task, but still attempted to submit the HIT. Additionally, we decided to implement an early exit in the workflow and treat the search task and multiple choice quiz as a prerequisite for our survey. This change more effectively filtered out workers who failed to faithfully attempt our task. Finally, we added an explicit uniqueness constraint to block workers from attempting our task multiple times.

Additionally, we received feedback from workers that our initial time limit (45 min) felt too rushed, leading us to increase the timer to 1 hour for our full study. However, we found that most workers spent less than the initial time limit on our task (averaging approximately 30 min). We paid workers $9.20 for the original expected work time of 45 minutes, the equivalent of the legal minimum wage in our state. We required that workers have more than 10,000 prior approved HITs with an approval rate greater than 98%.

4 DATA ANALYSIS

We collected a total of 540 responses from our main study (Group A: 202, Group B: 134, Group C: 201)[2]. We filtered out 81 responses (15%) during our preprocessing stage to account for workers who passed our initial quality control checks during the search task but recorded inattentive or careless responses in the subsequent survey. Following guidelines for identifying careless responses [6, 26, 37, 46], we analyzed response patterns and self-consistency.

Concretely, we filtered out responses that had (1) abnormally long unbroken strings (i.e., length > 8) of identical responses (e.g., a respondent answering a series of 18 consecutive questions with the same Likert-scale rating), (2) high overall numbers of inconsistent responses for positively and negatively keyed item pairs (i.e., total pairs > 4), and (3) high amounts (i.e., total > 5) of responses that were more than 2 points apart for highly similar question pairs. Additionally, we filtered out items from five aspects (i.e., *immediacy*, *efficiency*, *criticism*, *completeness*, *explanation fairness*) that produced inconsistent responses across all users. Unlike previous work that sometimes imposed even stricter criteria for preprocessing, we relaxed thresholds due to the length of our overall task and survey; we believe that workers who faithfully completed our task may not have any malicious intent or intentional carelessness, but instead that as workers see more questions in the survey, their chance likelihood of giving a single inconsistent answer should be accounted for. Additionally, we also checked for potential ordering effects in the response data and found none (Pearson's r=-0.012 between the position at which an item was presented and its response). A total of 459 valid responses were retained after filtering.

Best practices state the minimum sample size for Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) are 150 and 200, respectively [60, 67]. The 459 responses (Group A: 176, Group B: 110, Group C: 173) were randomly split into two sets: 200 responses were used for EFA and 259 responses were used for CFA. We used *factor_analyzer* and *semopy* [47] Python packages to conduct EFA and CFA, respectively[3].

4.1 Exploratory Factor Analysis (EFA)

The purpose of EFA is to determine the number of latent factors and the specific questionnaire items that measure each factor. This is done by examining the covariances in the observed data and grouping together items that are correlated into factors [60].

To ensure that our sample size of 200 was sufficiently large and was suitable for EFA, we followed the accepted practice of using Barlett's Test of Sphericity and Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy [67]. Barlett's Test determines if correlations between items are large enough for reduction by comparing the correlation matrix to the identity matrix. Statistically significant results indicate that the correlation matrix is not orthogonal and thus, the data is suitable for factor analysis. However, this test is sensitive to sample size, and it is recommended that additional evidence be provided to show factorability [60]. To supplement Barlett's Test, KMO measures the proportion of variance among variables that may be attributed to common variance to determine the adequacy of the sample size for factor analysis. Tabachnick and Fidell [60] recommend that results be greater than 0.6. Barlett's test resulted in a value of 11069.50 ($p < 0.001$) and KMO resulted in a value of 0.98. Both values indicate that our data was suitable for factor analysis [60, 67].

To extract factors, we performed factor analysis as its goal is to understand the latent constructs that contribute to the variance among observations, which is more suitable for latent variable detection and scale development over other extraction methods such as PCA [15, 67]. Specifically, we extracted factors using principal axis factoring (PAF), a least squares estimation of the latent factor model that minimizes the sum of the ordinary least squares [16, 17].

Since human behavior is rarely independent between functions, we assumed observations to be correlated and applied the appropriate oblique (promax) rotation to improve interpretability and clarify the factor solution structure by maximizing high item loadings and minimizing low item loadings [60, 66, 67]. Oblique rotation allows for inter-factor correlation, versus the alternative orthogonal rotation, which produces uncorrelated factors. When factors are entirely uncorrelated, both methods yield similar results [15].

To determine the number of factors to preserve, there are several popular factor retention methods such as a Scree test [12], Kaiser's criterion [38], or parallel analysis [34]. However, there is no clear consensus in the literature on which method is most reliable. Further, recent analysis shows that these "traditional" methods originating from the 1950s-1960s can fail to identify the best fitting model under certain conditions [4]. Kaiser's criterion suggests retaining those factors whose eigenvalues are greater than 1.0, which suggested we should retain 2 factors. Scree plot examination can often be unclear, since it calls for visual inspection to determine a "leveling off" point in the graph, and thus, its subjective nature can make the test unreliable. Parallel analysis involves calculating eigenvalues from a randomly generated dataset and comparing the values to the observed matrix, which indicated we should keep one factor (Figure 4). Overall, since no consensus was drawn from these methods, we tested both one and two factor solutions (i.e., first order and hierarchical two factor models) in Section 4.2.

We conducted an additional round of EFA fixing the number of factors to two and discarded items with weak factor loadings less than 0.4, cross-loading differences greater than 0.15, absolute loadings on multiple factors greater than 0.4, or weak communality h^2 of less than 0.4, as suggested by [14, 60, 67]. It is also important to note that at this stage, it is suggested to approximate a *simple structure* [61], meaning that factor groupings should seek to have intuitive meaning and items should only load on a single factor. To achieve this goal, researchers have suggested retaining at least three items per factor, deleting items that do not quite fit in with the rest of their factor grouping, or even repeating the study with additional items that are hypothesized to contribute to a specific factor [60, 67]. In other words, criteria for factor extraction and retention should not be interpreted as a strict rule, but instead,

[2] There is a slight imbalance despite conditions being randomly assigned, but distributions are roughly preserved across groups before and after preprocessing.
[3] For reproducibility purposes, we make all study code and data publicly available.

Table 2: EFA factor loadings

Factor 1	Factor 2	h^2	Questionnaire Item
0.94	0.21	0.93	**15.** This system would work well in a different search task (i.e. looking up medical papers to diagnose a patient).
0.84	-0.05	0.71	**37.** I would use this search engine in my everyday life.
0.83	-0.10	0.70	**21.** The results page provides me enough information to find the answers I am looking for effectively.
0.78	-0.15	0.62	**41.** The presentation of the results leads me to believe the results are ordered correctly.
0.76	-0.20	0.62	**49.** The results match my expectations and I agree with them.
0.75	-0.21	0.61	**47.** I trust that the results are ordered correctly and system will order results correctly for other queries.
0.54	-0.34	0.41	**19.** I can easily understand the contents of the results page.
0.13	0.93	0.88	**26.** If I change the query, I do not know how it will affect the result ordering.
-0.06	0.89	0.79	**0.** I do not understand why the results are ordered the way they are and would not be able to recreate the orderings myself.
-0.01	0.87	0.76	**6.** I think I need more information to understand why the given query produced the displayed results.
-0.06	0.87	0.76	**22.** I do not understand the document properties that cause some results to be ordered higher than others.
0.11	0.87	0.77	**12.** I am unable to see and understand how changes in the query affect the result ordering.
-0.16	0.78	0.63	**38.** The result interface does not help me understand the true decision making process of the search engine ranker.
-0.16	0.77	0.61	**24.** I do not understand why each result is ordered in a certain place.
-0.20	0.75	0.60	**4.** I'm unable to follow how the search engine ordered the results.
-0.16	0.74	0.57	**2.** It's difficult for me to break down each of the search engine's components and understand why the results are ordered the way they are.
-0.12	0.66	0.45	**8.** I do not know how confident the search engine is that its displayed orderings are correct.
-0.27	0.64	0.49	**46.** I do not trust that the results are ordered correctly and that the system will correctly order results for other queries.
-0.24	0.63	0.45	**34.** The format and amount of information provided in the result interface is not enough to help me understand why the results are ordered the way they are.

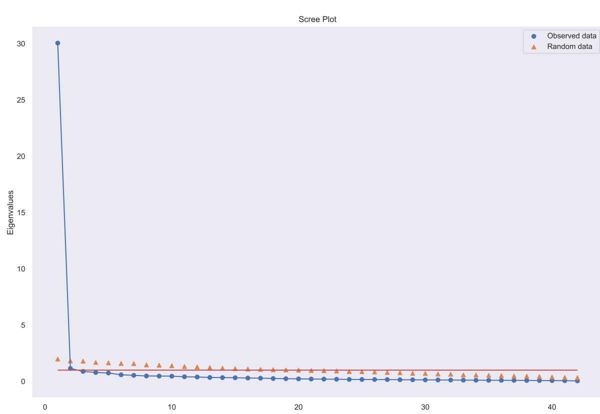

Figure 4: EFA scree plot

interpretability and other practical considerations should be taken into high account [4, 67].

A final round of EFA was conducted to ensure the factor solution was not greatly affected after item deletion and we analyzed the groupings for interpretability. Table 2 shows the final proposed item groupings and resulting factor loadings. Out of our original pool of 52, we retain twelve items in Factor Group 1 and seven items in Factor Group 2.

4.2 Confirmatory Factor Analysis (CFA)

While EFA is used to determine a candidate model structure, CFA is used to confirm the EFA-derived model fit. Goodness of fit is assessed by examining how closely the model-estimated covariance matrix aligns with the observed covariance matrix [10, 63] on a held-out set of data. SEM is commonly used to confirm the fit of potential model structures. To approximate the covariance matrix, we followed standard practice of using maximum likelihood (ML) estimation, which maximizes the likelihood that the model estimated parameters fit the observed data [17, 63].

We tested our EFA-derived hierarchical two-factor model on a held-out data set of 259 responses, which exceeded the recommended minimum size of 200 [45, 67]. Final standardized factor loadings [-1.00, 1.00] are reported in Table 3, in addition to the standard errors associated with each factor loading.

We also compared our proposed hierarchical two-factor model to a null model, where all items were assumed to be independent (covariances were fixed to 0), and a first-order factor model, where all items loaded onto a single latent factor of explainability. In Figure 5, we show a visual representation of our hierarchical factor model. In SEM, *path diagrams* can be used to represent relationships between latent factors and items. Factors appear in circles or ovals, and items appear in squares or rectangles. Directed arrows connect entities in a path diagram and can either be single-or double-headed. A single-headed arrow indicates a direct relationship between two variables where the variable with the arrow pointing to it is said to *load* on the other, and edge weight represents a regression coefficient. A double-headed arrow also indicates a relationship between variables, but with no direction of effect, and edge weight simply represents covariance.

Table 3: Final factor loadings

Factors & Items	Original Factor Label	Std. Loading	Std. Error
Factor 1		0.997	0.000
41	Plausibility	0.905	0.096
49	Justifiability	0.888	0.097
47	Trustworthiness	0.878	0.096
21	Informativeness	0.868	0.094
37	Acceptability	0.851	0.110
19	Understandability	0.825	0.087
15	Transferability	0.703	0.000
Factor 2		-0.944	0.065
2	Decomposability	0.938	0.094
22	Global Interpretability	0.928	0.098
24	Local Interpretability	0.921	0.098
0	Simultability	0.911	0.099
38	Faithfulness	0.894	0.097
4	Algorithmic Transparency	0.884	0.097
46	Trustworthiness	0.877	0.091
6	Causaility	0.842	0.095
8	Uncertainty	0.792	0.092
34	Units of Explanation	0.789	0.094
12	Visibility	0.723	0.092
26	Counterfactuals	0.704	0.000

Common research practice in SEM is to use a chi-square goodness of fit test, however, due to its sensitivity to sample size, it is suggested to supplement this test with alternative fit indices [7, 35, 41]. Some researchers attempt to minimize the impact of sample size by reporting a relative chi-square statistic (χ^2 / df), but there is no strong consensus on acceptable ratios, as values vary from 5.0 [65] to 2.0 [60]. We report the results of both chi-square statistics for completeness, but followed standard practice of assessing model fit by examining two additional categories of fit indices: absolute fit to measure how well our model fit the observed data and incremental fit to measure our proposed model against a baseline model [33, 35, 41]. Table 4 shows that our hypothesized hierarchical two-factor model achieved a better fit over the null and first-order models, and is well within the acceptable ranges for all fit statistics. For the incremental fit indices, we report the Comparative Fit Index (CFI) and the Non-Normed fit index (NNFI), also known as the Tucker Lewis Index (TLI). The CFI was 0.968 and the NNFI was 0.964, both above the standard acceptable value of 0.95 [35]. For absolute fit indices, we supplemented the chi-square test with the root mean square error of approximation (RMSEA) and the standardized root mean square residual (SRMR), with values

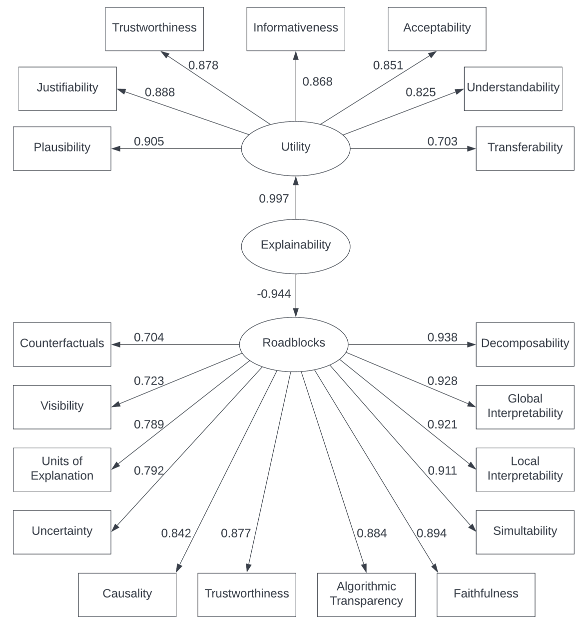

Figure 5: *Path diagram* for proposed structural equation model for modeling explainability.

at 0.068 and 0.021, respectively, which were below the acceptable levels of 0.07 and 0.08 [35, 59].

5 DISCUSSION

5.1 Dimensions of Search Explainability

Here, we present exactly which aspects respondents found important for explainability. The results of CFA confirm that items organize into two distinct factor groups (Table 3). Examining the items in each factor reveals that constructs are separated into positive and negative traits. In Factor 1, we retain seven items referring to positive attributes and in Factor 2, we retain twelve items referring to negative attributes. While there is no ideal label for these overarching concepts as they are composites of multiple aspects, we attribute names that we feel encompass the nature of the items in the respective groups for a clearer, more targeted discussion of properties. Thus, we define the factor groups as **(A)** Factor 1: *utility* and **(B)** Factor 2: *roadblocks*. The first factor broadly corresponds to the system's *utility*, and falls in line with existing evaluation strategies for explainability, where researchers measure the explainability of their system by testing its usefulness within some context or application [44, 55]. The second factor compiles a range of critical *roadblocks* that can be thought of as properties a system might lack in order to be fully explainable. In Table 3, we also show the original aspect labels associated with each item, derived from our literature review during the questionnaire development phase. Intuitively, item-factor loadings represent how significant an item is to overall factor. For example, Item 41 (*plausibility*) is a strong indicator of Factor 1 (*utility*). In other words, we can interpret this as the higher the *plausibility* score, the more useful its explanations

Table 4: Global fit statistics for CFA

Model	$\chi^2 \downarrow$	$df \downarrow$	χ^2 / $df \downarrow$	CFI ↑	NNFI ↑	RMSEA ↓	SRMR ↓
Null model	5792.80	171	33.88	–	–	–	–
First-order model	465.08	152	3.06	0.937	0.944	0.089	0.026
Hierarchical two-factor model	**327.55**	**150**	**2.18**	**0.968**	**0.964**	**0.068**	**0.021**

are. Conversely, Item 2 (*decomposability*) is a strong indicator of Factor 2 (*roadblocks*). Since Factor 2 has a negative loading, higher responses on Item 2 indicate a strong *lack of decomposability* and lower explainable capability.

Although our literature review produced 26 potential facets contributing to the overall notion of explainability, we find that users are only concerned with 18 of them. While we cannot draw conclusions from the 5 factors that were filtered out during preprocessing due to highly inconsistent responses across all users, we find that users discount 3 aspects entirely: *compositionality* [21, 22, 42], *model fairness* [3, 43], and *accuracy* [36]. We note that our findings are pertinent only to Web search systems, and that conducting this study in other search domains (i.e., clinical abstract search) or other ML tasks (i.e., image classification) may find these aspects to be important to explainability (more in Section 5.3).

Nonetheless, our results support the multidimensionality of explainability posited by recent literature [22, 43, 48], and we further this growing body of work by contributing empirical evidence that these factors group between positive and negative facets that describe the *utility* and *roadblocks to explainability* of search systems. While this strong split among positive and negative factors is not seen in other psychometric multidimensional modeling work in IR, such as Zhang et al.'s work [69] to model multidimensional relevance, researchers posit that modeling negative aspects of user experience and reasons for non-use can be valuable in future system design [5, 28, 40, 49].

5.2 Implications for Design and Evaluation

The dimensions of explainability we identify can be used to seamlessly integrate explainable features into existing search systems. While it is important to design systems that will instill trust in AI systems in an age of misinformation, it is also our responsibility that our systems have minimal negative impact on the search experience. To do so, we must acknowledge the cost and risk of designing and integrating explainable systems. No matter how theoretically sound a new feature is, there is no guarantee that all users will be satisfied with it because they have grown accustomed to the existing system. In some cases, unintuitive changes may even hinder the user experience as they may increase cognitive load and require users to relearn system functions. Thus, since Web search systems attract millions of users per day, we must take into consideration the functionality and usability of novel explainable elements.

This is not a trivial task; one goal of adaptive user interface design in HCI research is to examine how incremental changes can be made with minimal disruption to the user experience. Todi et al. [62] note that *estimation of utility* is fundamental to designing new features, yet is notoriously difficult to estimate. Our work is a step towards deploying explainable interfaces that cause minimal disruption to the user experience by introducing our model and questionnaire as an evaluation framework that will enable researchers to make targeted system improvements and compare their system against others.

An ideal, perfectly explainable system will then receive higher responses on items in the *utility* factor and lower responses on items in the *roadblocks* factor to maximize the overall explainability score. In practice, we imagine that real systems may produce varying response levels and our factor model can aid in targeted design improvements, e.g., via A/B testing. Moreover, lower scores on facets relating to *utility* and higher scores on facets regarding *roadblocks* indicate areas for improvement and the specific aspects can directly be inferred from the response data.

5.3 Limitations

One potential limitation of our study is that due to survey-time constraints, the search interfaces presented to users were not entirely dynamic. In the initial search task, we presented users with a mock system with pre-populated query and results pages, though users were still able to interact with the system by navigating through the pages and clicking on links. However, the mock system presented was designed to resemble commercial search engines that we expect most users to use on an everyday basis and the first portion of our study served mostly as a form of quality control in addition to preparing participants to complete our survey. In the follow-up survey, we presented users with a static mock interface. As a result, we had to discard certain items from our initial questionnaire since the questions included references to features users did not have access to in a static version. Although it is possible that the static presentation affected user responses, we find the resulting factor groupings to be interpretable and intuitive.

Though some researchers criticize the quality of data collected via crowdsourcing due to poor compensation and advocate for automated evaluation [30, 32], choosing a proxy evaluation method can be challenging [21] and results do not capture the true feelings of end users who ultimately, will be using these applications. Since the goal of explainability is to provide insight into a model's decision-making process in *human-understandable* terms, it naturally follows that we should assess these systems with humans to provide more impactful evaluations.

Our study employs stringent but necessary controls in order to collect faithful responses and mitigate concerns of low-quality data. While it is possible such controls could introduce a bias towards demographics that are more willing to complete a longer task, Difallah et al. [19] analyzed worker propensity with demographic correlation and found that "most demographic variables are not affected by [such] selection biases" (with the exception that Indian workers may be overrepresented in the pool). MTurk does not release demographic information about workers nor did we set such selection criteria, choosing to follow common practice of selecting workers based on fidelity (e.g., HIT approval rates, # of completed HITs) to improve the chance of receiving high-quality data [51]. However, we note that HIT batches were released in the AM (ET) and most submissions were completed by EOD or very early AM the following day, so it is possible that most responses were collected from time zones whose daytime overlaps the most with the collection time frame (i.e., US/Europe).

Additionally, it is important to note that explainability is highly domain-specific and can change depending on the intended user and task. Users search to satisfy a strongly context dependent information need. For example, a clinician using a diagnostic decision support system may require more detailed information and prefer certain system attributes over a journalist fact-checking sources. Thus, we make the distinction that the dimensions we find pertinent to explainability in this paper are limited to Web search systems used by the everyday layperson and may not hold true for all conceivable domains. However, the methodology outlined in this work can be easily adapted to create multidimensional models of explainability for other domains and ML tasks.

However, it is important to acknowledge the potential risks of modeling error in other domains and among aspects. For example, the cost of error in the biomedical domain may be much higher than the cost of error in the everyday search use case we examine in this work. Additionally, getting aspects such as *trustworthiness* or *uncertainty* is potentially more risky than getting *visibility* wrong.

6 CONCLUSION

In this paper, we establish a user-centric definition of search system explainability grounded in recent literature. Based on a large-scale crowdsourced user study and factor analysis, we show that users consider both *utility* and critical *roadblock* factors in explainable search systems. The resulting factor model will not only allow for a direct comparison between explainable systems, but will also enable informed trade-offs between system quality and explainability. In the future, we plan to further assess the validity of our proposed factor model as an evaluation tool through additional case studies. The methodology introduced in this work also has the potential to be applied to other IR domains, and the wider NLP and ML communities.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. In *Ethics of Data and Analytics*. Auerbach Publications, 254–264.

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[4] Max Auerswald and Morten Moshagen. 2019. How to determine the number of factors to retain in exploratory factor analysis: A comparison of extraction methods under realistic conditions. *Psychological methods* 24, 4 (2019), 468.

[5] Eric PS Baumer, Jenna Burrell, Morgan G Ames, Jed R Brubaker, and Paul Dourish. 2015. On the importance and implications of studying technology non-use. *interactions* 22, 2 (2015), 52–56.

[6] Tara S Behrend, David J Sharek, Adam W Meade, and Eric N Wiebe. 2011. The viability of crowdsourcing for survey research. *Behavior research methods* 43, 3 (2011), 800–813.

[7] Peter M Bentler and Douglas G Bonett. 1980. Significance tests and goodness of fit in the analysis of covariance structures. *Psychological bulletin* 88, 3 (1980), 588.

[8] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. 2021. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 401–413.

[9] Adrien Bibal and Benoît Frénay. 2016. Interpretability of machine learning models and representations: an introduction.. In *ESANN*.

[10] Timothy A Brown and Michael T Moore. 2012. Confirmatory factor analysis. *Handbook of structural equation modeling* 361 (2012), 379.

[11] Nadia Burkart and Marco F Huber. 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research* 70 (2021), 245–317.

[12] Raymond B Cattell. 1966. The scree test for the number of factors. *Multivariate behavioral research* 1, 2 (1966), 245–276.

[13] Daniel Cohen, Bhaskar Mitra, Oleg Lesota, Navid Rekabsaz, and Carsten Eickhoff. 2021. Not all relevance scores are equal: Efficient uncertainty and calibration modeling for deep retrieval models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 654–664.

[14] Andrew L Comrey and Howard B Lee. 2013. *A first course in factor analysis*. Psychology press.

[15] Anna B Costello and Jason Osborne. 2005. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research, and evaluation* 10, 1 (2005), 7.

[16] Robert Cudeck and Robert C MacCallum. 2007. *Factor analysis at 100: Historical developments and future directions*. Routledge.

[17] Joost CF De Winter and Dimitra Dodou. 2012. Factor recovery by principal axis factoring and maximum likelihood factor analysis as a function of factor pattern and sample size. *Journal of applied statistics* 39, 4 (2012), 695–710.

[18] Cecilia di Sciascio, Vedran Sabol, and Eduardo Veas. 2015. uRank: Visual analytics approach for search result exploration. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE Computer Society, 217–218.

[19] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and dynamics of mechanical turk workers. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 135–143.

[20] Derek Doran, Sarah Schulz, and Tarek R Besold. 2017. What does explainable AI really mean? A new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794* (2017).

[21] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[22] Finale Doshi-Velez and Been Kim. 2018. Considerations for evaluation and generalization in interpretable machine learning. In *Explainable and interpretable models in computer vision and machine learning*. Springer, 3–17.

[23] Mike Furr. 2011. *Scale construction and psychometrics for social and personality psychology*. SAGE publications ltd.

[24] R Michael Furr. 2021. *Psychometrics: an introduction*. SAGE publications.

[25] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.

[26] Lewis R Goldberg and James M Kilkowski. 1985. The prediction of semantic consistency in self-descriptions: Characteristics of persons and of terms that affect the consistency of responses to synonym and antonym pairs. *Journal of personality and social psychology* 48, 1 (1985), 82.

[27] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine* 38, 3 (2017), 50–57.

[28] Howard Greisdorf. 2003. Relevance thresholds: a multi-stage predictive model of how users evaluate information. *Information Processing & Management* 39, 3 (2003), 403–423.

[29] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense advanced research projects agency (DARPA), nd Web* 2, 2 (2017), 1.

[30] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P Bigham. 2018. A data-driven analysis of workers' earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–14.

[31] Marti A Hearst. 1995. Tilebars: Visualization of term distribution information in full text information access. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 59–66.

[32] Bernease Herman. 2017. The promise and peril of human evaluation for model interpretability. *arXiv preprint arXiv:1711.07414* (2017).

[33] Daire Hooper, Joseph Coughlan, and Michael R Mullen. 2008. Structural equation modelling: Guidelines for determining model fit. *Electronic journal of business research methods* 6, 1 (2008), pp53–60.

[34] John L Horn. 1965. A rationale and test for the number of factors in factor analysis. *Psychometrika* 30, 2 (1965), 179–185.

[35] Li-tze Hu and Peter M Bentler. 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal* 6, 1 (1999), 1–55.

[36] Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685* (2020).

[37] John A Johnson. 2005. Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of research in personality* 39, 1 (2005), 103–129.

[38] Henry F Kaiser. 1960. The application of electronic computers to factor analysis. *Educational and psychological measurement* 20, 1 (1960), 141–151.

[39] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems* 29 (2016).

[40] Hyeji Kim, Inchan Jung, and Youn-kyung Lim. 2022. Understanding the Negative Aspects of User Experience in Human-likeness of Voice-based Conversational Agents. In *Designing Interactive Systems Conference*. 1418–1427.

[41] Theresa Kline. 2005. *Psychological testing: A practical approach to design and evaluation*. Sage.

[42] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. 2019. Human evaluation of models built for interpretability. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 59–67.

[43] Zachary C Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[44] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[45] Robert C MacCallum, Keith F Widaman, Shaobo Zhang, and Sehee Hong. 1999. Sample size in factor analysis. *Psychological methods* 4, 1 (1999), 84.

[46] Adam W Meade and S Bartholomew Craig. 2012. Identifying careless responses in survey data. *Psychological methods* 17, 3 (2012), 437.

[47] Georgy Meshcheryakov, Anna A Igolkina, and Maria G Samsonova. 2021. semopy 2: A structural equation modeling package with random effects in python. *arXiv preprint arXiv:2106.01140* (2021).

[48] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2022. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *arXiv preprint arXiv:2201.08164* (2022).

[49] Abu Saleh Md Noman, Sanchari Das, and Sameer Patil. 2019. Techies against Facebook: understanding negative sentiment toward Facebook via user generated content. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–15.

[50] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[51] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. 2014. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior research methods* 46, 4 (2014), 1023–1031.

[52] Sayantan Polley, Rashmi Raju Koparde, Akshaya Bindu Gowri, Maneendra Perera, and Andreas Nuernberger. 2021. Towards trustworthiness in the context of explainable search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2580–2584.

[53] Pernilla Qvarfordt, Gene Golovchinsky, Tony Dunnigan, and Elena Agapie. 2013. Looking ahead: Query preview in exploratory search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 243–252.

[54] Jerome Ramos and Carsten Eickhoff. 2020. Search result explanations improve efficiency and trust. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1597–1600.

[55] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[56] Tobias Schnabel, Saleema Amershi, Paul N Bennett, Peter Bailey, and Thorsten Joachims. 2020. The Impact of More Transparent Interfaces on Behavior in Personalized Recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 991–1000.

[57] Jaspreet Singh and Avishek Anand. 2019. Exs: Explainable search using local model agnostic interpretability. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 770–773.

[58] Dylan Slack, Sorelle A Friedler, Carlos Scheidegger, and Chitradeep Dutta Roy. 2019. Assessing the local interpretability of machine learning models. *arXiv preprint arXiv:1902.03501* (2019).

[59] James H Steiger. 2007. Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual differences* 42, 5 (2007), 893–898.

[60] Barbara G Tabachnick, Linda S Fidell, and Jodie B Ullman. 2007. *Using multivariate statistics*. Vol. 5. pearson Boston, MA.

[61] Louis Leon Thurstone. 1947. Multiple-factor analysis; a development and expansion of The Vectors of Mind. (1947).

[62] Kashyap Todi, Gilles Bailly, Luis Leiva, and Antti Oulasvirta. 2021. Adapting user interfaces with model-based reinforcement learning. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[63] Jodie B Ullman and Peter M Bentler. 2012. Structural equation modeling. *Handbook of Psychology, Second Edition* 2 (2012).

[64] Ellen Voorhees. 2005. Overview of the TREC 2004 Robust Retrieval Track. https://doi.org/10.6028/NIST.SP.500-261

[65] Blair Wheaton, Bengt Muthen, Duane F Alwin, and Gene F Summers. 1977. Assessing reliability and stability in panel models. *Sociological methodology* 8 (1977), 84–136.

[66] Brett Williams, Andrys Onsman, and Ted Brown. 2010. Exploratory factor analysis: A five-step guide for novices. *Australasian journal of paramedicine* 8, 3 (2010).

[67] Roger L Worthington and Tiffany A Whittaker. 2006. Scale development research: A content analysis and recommendations for best practices. *The counseling psychologist* 34, 6 (2006), 806–838.

[68] Yunjie Xu and Zhiwei Chen. 2006. Relevance judgment: What do information users consider beyond topicality? *Journal of the American Society for Information Science and Technology* 57, 7 (2006), 961–973.

[69] Yinglong Zhang, Jin Zhang, Matthew Lease, and Jacek Gwizdka. 2014. Multidimensional relevance modeling via psychometrics and crowdsourcing. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 435–444.