Multidimensional Relevance Modeling via Psychometrics and Crowdsourcing

Yinglong Zhang[1], Jin Zhang[2], Matthew Lease[1], and Jacek Gwizdka[1]
School of Information[1], Department of Integrative Biology[2]
University of Texas at Austin
{ylzhang,zj,ml}@utexas.edu, sigir2014@gwizdka.com

ABSTRACT
While many multidimensional models of relevance have been posited, prior studies have been largely exploratory rather than confirmatory. Lacking a methodological framework to quantify the relationships among factors or measure model fit to observed data, many past models could not be empirically tested or falsified. To enable more positivist experimentation, Xu and Chen [77] proposed a psychometric framework for multidimensional relevance modeling. However, we show their framework exhibits several methodological limitations which could call into question the validity of findings drawn from it. In this work, we identify and address these limitations, scale their methodology via crowdsourcing, and describe quality control methods from psychometrics which stand to benefit crowdsourcing IR studies in general. Methodology we describe for relevance judging is expected to benefit both human-centered and systems-centered IR.

General Terms
Experimentation, Human Factors, Measurement, Reliability

Keywords
Relevance judgment; psychometrics; crowdsourcing.

1. INTRODUCTION
Despite relevance being 80 years old [13] and many attempts to define its contributing factors (for review see [11, 51, 61]), no conclusive results have been drawn and debate continues [38]. Xu and Chen [77] lamented in 2006 how little we still know about the factors influencing relevance judgments, writing:

> "...there is no agreement on factors beyond topicality, neither in terms of what they should be nor of how important they are… Naturalistic inquiry with qualitative research methods has been advocated and adopted by many researchers… [yet] almost no study of relevance judgment had adopted a confirmatory approach."

To address this, they proposed a novel statistical framework based upon *psychometrics* [14] for modeling relevance as a function of any number of contributing factors. In so doing, they sought to enable a new thrust of positivist relevance studies in which a solid foundation statistical hypothesis testing would offer new traction on this old and thorny issue. While prior studies had posited a wide range of alternative factors with little resolution, factors could now be integrated and empirically assessed to determine the relative impact of each upon overall relevance. Moreover, the extensible framework permitted any newly hypothesized factors to be similarly incorporated and analyzed in order to test if their inclusion would enable the model to better explain observed data.

While the overall framework and goals of Xu and Chen's work continue to offer enduring value today, our review of the actual mechanics of their psychometric approach has revealed several methodological concerns which could threaten the validity of results derived from their framework as originally proposed. The primary contribution of our paper is to delineate and rectify these methodological limitations such that their framework can live up to its full potential. As an innovative aspect of their work was introducing IR to *psychometrics* methodology, we expect that the novelty of psychometrics has contributed to limited adoption of their ideas. We provide a brief primer to help remedy this.

Another contribution of our work is adapting this approach from a traditional interactive research design using student participants to a systems-oriented relevance judging approach using crowdsourced data collection. As such, our work straddles the traditional divide between user-centered and systems-centered IR research. Better understanding the factors influencing relevance decisions has potential to not only better explain end-user behavior and expectations, but also offer new insights into oft-reported disagreement in systems-oriented relevance judging [3, 40, 44, 72]. Potential benefits include better understanding: 1) the type and importance of varying relevance criteria; 2) where search systems might best focus effort beyond topicality; 3) how multidimensional judging may yield more reliable overall relevance judgments; and 4) how multidimensional judgments can be effectively collected at scale to enable future systems-oriented evaluations beyond traditional Cranfield topicality [57].

Crowdsourced collection of subjective data is now firmly-established in the behavioral sciences, having been shown to faithfully reproduce many past findings [9]. As such, best practices for crowdsourcing from psychometrics may benefit not only user-centered IR studies, by increasing sample size and diversity, but also evaluation of IR systems, by increasing quality, dimensionality, and diversity of judgments. We might evaluate system effectiveness over multiple, weighted relevance dimensions, or maximize diversity metrics over a distribution of subjective judgments for the same topical intent. We discuss well-established survey design techniques for ensuring data validity with crowdsourcing. We also posit that relevance judging might be more reliably crowdsourced by collecting multi-dimensional judgments, then aggregating across dimensions at the individual level. For reproducibility, our study data can be obtained online[1].

SIGIR '14, July 06 - 11 2014, Gold Coast, QLD, Australia.

ACM 978-1-4503-2257-7/14/07…$15.00.
http://dx.doi.org/10.1145/2600428.2609577.

[1] www.ischool.utexas.edu/~ml/data/zhang-sigir14-data.zip

2. RELATED WORK

Relevance definitions [51, 61, 63, 73] show a long-standing dichotomy between objective vs. subjective perspectives [7, 11, 38]. A growing shift from objective algorithm-oriented relevance to subjective relevance [11, 60, 62] is emphasized by work inferring neuro-physiological relevance [31, 52]. In contrast with these experimentally advanced but not easily scalable approaches, a psychometrics approach can be crowdsourced.

2.1 Relevance Criteria

Since relevance is subjective, what are the factors that contribute to relevance judgments? We have known for decades that topicality [34, 51] alone is not sufficient [5, 12]. What other criteria are considered by users? Early in 1960s, Rees and Schultz [58] suggested 40 variables related to relevance, Cuadra and Katter [21] found 38 factors contributing to relevance, Cool et al. [19] identified six facets of relevance judgments, and Taylor et al., [68] found that relevance criteria change across search task stages.

The list of proposed relevance judgments criteria is much longer, and the plurality of these proposals is problematic and carries a number of limitations. Firstly, there are typically many factors in each relevance model. It may be virtually impossible to ask users to assess all of these factors. Secondly, different studies seem to use synonyms or near-synonyms for the same criteria (i.e. utility and usefulness [29]). Thirdly, some factors overlap in their meaning (i.e. new, novel and recent). Fourthly, often there is no distinction made between the effects of IR systems and documents. For example, the accessibility of a document [62] is a part of efficiency of access [11] and a property of the IR system. The relevance of the document should be defined by properties of the document itself, not the IR system. Past work to address some of the above-mentioned problems include the work of Barry and Schamber [7], who compared their own studies conducted in two contexts (academic and weather IR), found a significant overlap of criteria. Bateman [8] suggested that the major criteria were fairly stable across different situations though the set of criteria might change. More recently, Saracevic [62] in his synthesis of several decades of work proposes seven groups of relevance criteria: content, objects, validity, situational match, cognitive match, affective match, belief match. Despite such comprehensive attempts at improving our understanding of relevance criteria, establishing them still remains a work-in-progress [38].

Most previous studies of relevance criteria relied on data collected from questionnaires in which items were created based on a combination of intuition and prior research. In a few cases (e.g., [4]), criteria established in prior research were verified with users' comments (e.g., from a concurrent or retrospective talk aloud protocol). Generally, there is a lack of approaches that ground subjective user responses in a theoretical framework.

Building upon Grice's framework of human communication [30], Xu and Chen [77] focus upon the five relevance criteria below:

Topicality. Rooted in the heart of human document indexing [49], topicality combines aboutness [49], relatedness [67], and topical relevance [28, 61]. Boyce [12] indicates that users first judge the topicality of document, and only then consider other factors.

Novelty. Often, to be considered useful, documents must inform the user beyond what they already know [43]. Studies by Harter [32] suggest that citations already known to users were not considered relevant due to not yielding a change cognitive state.

Understandability. A document must be understood to be useful. Studies show across users' expertise levels that use of unfamiliar jargon significantly reduces determination of relevance [22, 67].

Scope addresses breadth and depth [69]. Levitin and Redman [45] argue that scope is important as it affects perception of document quality. Barry equates depth/scope and gives examples [6]. Intuitively, a user wants a document broad and specific enough to satisfy the given information need, yet without being so broad or specific that desired information is difficult to extract.

Reliability. Typically, information must be perceived as accurate to be considered relevant. Petty and Cacioppo [15] suggest that message receivers evaluate the quality of the message before accepting it. Reliability was considered key to relevance [39].

While we adopt these same five criteria in our study, prior findings [8, 59, 67, 68, 71] support that people's criteria to relevance judgment change in different tasks, given different cognitive states as a task changes or the stages of a task are switched [62]. The methodology we propose for assessing relevance criteria is independent of the criteria chosen and search task, and it is critical to distinguish any choice of hypothesized multidimensional factors from how such a hypothesis is tested.

We believe the primary contribution of both Xu and Chen's work and our own is the methodology for positivist hypothesis testing rather than the actual findings regarding relevance. One may begin with a theory-driven hypothesis to test, or in data-driven hypothesis generation, empirically posit a set of factors, let the data speak for itself, and then seek an explanatory theory. Our decision to use the same five relevance criteria was intended to make our two studies maximally similar so that any differences in methodology could be easily discerned by the reader. Secondarily, by preserving the same criteria across studies, we could most easily investigate whether the revised methodology we propose might lead to real differences in findings for these same criteria.

2.2 Crowdsourcing

A variety of recent work has investigated crowdsourcing methods for IR data collection [1], both for Cranfield-style relevance judging [35] and interactive IR studies [78]**.** Potential benefits include faster, easier, cheaper, and scalable data collection, increasing diversity of data available beyond that provided by traditional assessors and university students, and the potential greater similarity between the crowd and typical IR system users. The primary challenge of crowdsourcing is ensuring data quality while not incurring greater cost outweighing the actual benefits [10]**.** While a variety of platform and incentive models exist, research has predominantly focused on the pay-based Amazon Mechanical Turk platform (AMT), which we adopt in this work.

Typical quality control techniques include using gold-questions with known answers, and/or assigning the same task to multiple people and comparing agreement between their responses [35]. In this way, one can both assess whether individuals produced expected objective responses, and aggregate their responses to improve label quality. What these techniques largely fail to achieve, however, is quality control for subjective tasks. In fact, aggregation in such cases may actually eliminate valid diversity in responses that would be valuable to detect and model. Moreover, initial qualification tests do not verify ongoing behavior, and testable captchas are easily distinguishable from subjective questions. Turkers have developed their own lingo of "ACs" (attention checks) and "MCs" (memory checks) to describe such easily identifiable quality controls vs. actual task work [50].

An early proposed technique from the HCI community was to design tasks to be sufficiently effortful that it is no easier to cheat than to complete an assigned task in good faith [42, 46]. Similarly, while one can try to make tasks engaging and fun [24]**,** we would

prefer a methodology that does not require that work be made entertaining in order to be considered reliable. More systematic procedures, however, come from traditional survey design methodology which predates crowdsourcing. For example, we might pose nearly the same question twice, using slightly different wording, or negating the question when repeating it. For Likert scale questions, we can also check for constant neutral or near-neutral responses, which would pass the above checks. While there was early concern of data validity with behavioral crowdsourcing studies, mounting evidence has shown consistency of findings with those yielded by traditional lab studies [9].

2.4 Psychometrics and SEM

Psychometrics is the theory and methodology of psychological measurement of cognitive properties, covering techniques for both data collection and analysis [14]. Because cognitive traits are typically latent and cannot be measured directly [14, 55], one investigates the interrelationships between observed (i.e., measurable) *manifest variables* from which properties of *latent variables* can be indirectly deduced [14]. Because observed data are expected to exhibit substantial measurement error, multiple interrogation techniques are typically applied with repetition [14].

Within psychometrics, structural equation modeling (SEM) [36] provides a well-established framework for modeling latent factors, their inter-relationships, and relationships to observed data. SEM derives from path analysis, invented by Sewall Wright in 1921 [76]. Social and behavioral sciences begin using SEM in the early 1970s, where it has since become widely adopted. Under SEM, latent factors may be hypothesized *a priori* and/or emerge from the data through analysis. SEMs are defined by a set of equations between variables which must be solved from observed data. Linear models are common but not a limitation of SEM.

Like graphical models, SEMs can be fully defined by an equivalent graphical representation: the *path diagram* (e.g., see **Figures 2-3**). Murphy briefly reviews SEM vs. graphical models [54]. Following path analysis notation [76], observed variables are shown in SEM by boxes, while circles depict latent factors. Directed edges (i.e., *factor loadings*) denote causal relationships in the pointed direction (regression coefficients). A pointed-to variable is said to *load* on the factor pointing to it. Bi-directional edges denote correlation without causal interpretation. Edge weights for each case denote regression coefficients and covariance, respectively. Because we do not expect the model to perfectly explain observed data, a latent residual error term is typically associated with each observation and estimated with other model parameters. This latent error term may also be depicted by a circle, or omitted and implicitly assumed. A model is completely parameterized by its factor loadings, factor variances or covariances, and the residual error terms.

SEM begins with *Exploratory Factor Analysis* (EFA), in which statistical analysis proceeds without prior assumptions about the number of latent factors and relationships between latent factors and observed data (though we may have prior hypotheses). Model structure can be learned entirely from data, though the number of factors can also be individually fixed to impose a particular independence assumption. Starting from some initial maximal number of possible factors, we first assume all variables load on all factors. Statistical analysis is then employed to deduce both the number of factors to be kept and their associated edge weights. Those edge weights which are sufficiently low are then pruned from the model's structure to reduce model complexity.

Once the model's structure is determined, *Confirmatory Factor Analysis* (CFA) is employed to assess that particular structure, re-estimating parameters and testing model fit as a function of data likelihood. Positivist significance testing enables a proposed model to be rejected for failing to explain observed data with sufficient likelihood. Since a model cannot be proven correct, but only falsified, alternative hypotheses (i.e., competing models) are typically compared relative to one another. While we can evaluate competing models which encode alternative causation vs. correlation hypotheses, an experimenter must still proceed with care in distinguishing correlation vs. causation relationships.

Observed data are typically assumed to be continuous and normally distributed. Simulation studies have shown that typical model sizes can be estimated via maximum likelihood with only about 200 observed instances [64]. Studies have shown that ordinal categorical data can still be accurately modeled as continuous, provided there are at least five categories and approximate normality, as in Xu and Chen's study [77] and ours.

3. STUDY DESIGN

Our psychometric methodology includes: data collection, modeling dimensions of relevance and their relationships, and inferring the significance of each dimension for overall relevance. Rather than presuppose any particular definition of relevance, or posing any direct questions about this often tacit concept, we induce relevance as a latent variable. It is established based upon how well its inclusion better explains observed data. A strength of psychometric modeling is that it permits complex, latent factors to be robustly induced indirectly, as a data-driven, hierarchical combination of simpler factors which are more easily queried.

3.1 Survey Design

The first step of psychometric analysis is to design a questionnaire (known as the "instrument") which is issued to participants as a survey. One new to conducting surveys might assume this is as simple as listing one's questions, and many naïve crowdsourcing studies appear to do just that; when collected data (predictably) turns out to be poor, the researcher simply blames "spammers". In contrast, it is well-known in social sciences, marketing, and human-computer interaction fields that effective survey design is a science, and how you ask a question strongly influences the quality of the ultimate answer you receive [17].

The goal of our questionnaire is to measure each of the relevance criteria (see Section 2): topicality, novelty, understandability, scope, and reliability. Following best practices, each dimension is assessed using multiple questions (called "items" in psychometrics). We adopt a seven-point Likert scale with anchors at: 1 (strongly disagree), 4 (neutral), and 7 (strongly agree). Items were created following established principles [26]:

1. Content must reflect the intended psychological variable
2. Be straightforward and avoid complicated terms
3. Avoid leading or presumptive wording
4. Score scales should be "balanced" by including positively keyed and negatively keyed items.

The final principle above suggests that the overall questionnaire should be positively or negatively "keyed" so that agreement and disagreement can be expected roughly equally. Not only does a balanced distribution help keep respondents cognitively engaged, and avoid skewing results by skewed question polarity, but it also enables a method of quality control, as discussed below. It is not necessary that every factor have equally balanced questions.

Self-consistency is assessed by posing redundant pairs of highly-similar questions which each articulate the same underlying query with slightly different wording [9]. For example, "I think the information in this passage is wrong" and "I think some or all of the information in this passage is incorrect." In this case, we expect similar answers to each question pair and test for this. However, if all questions were similarly keyed, we could not detect an improper "constant" respondent who answered all such question-pairs positively or negatively. We thus also pose redundant pairs of oppositely-keyed questions. For instance, "It's easy for me to understand most of the information in this passage" and "It's difficult for me to understand most of the information in this passage". When analyzing responses, the scale of negatively keyed question is reversed, then responses are tested for similarity as with the highly-similar question pairs above.

Another intentional aspect of the above quality control design is making the task sufficiently effortful that it is no easier to "cheat" than to answer the questionnaire in good faith [42]. To some degree, one might regard our posing multiple questions for each relevance dimension to be similar to the common crowdsourcing practice of posing the same question to multiple respondents and checking for agreement (i.e., "plurality") [35]. However, testing self-consistency of individual respondents supports subjective data collection for tasks in which respondents cannot be expected to agree with one another or any fixed gold-standard.

A related question is how many questions to include in the survey? More questions could be more informative, but fatigue participants and increase time and cost of data collection. *Kenny's Rule* offers an established rule-of-thumb for determining how many questions to use per factor for modeling: "two might be fine, three is better, four is best, and anything more is gravy" [76]. However, because some of our questions likely will not correlate as closely with factors as intended, standard practice is to over-generate questions, expecting some will be later pruned during cognitive interviewing, pilot testing, and EFA analysis (further discussed below). For example, while EFA analysis permits more than 3 questions to be retained, we keep only 3 questions given strength of factor loadings and Kenny's Rule above.

Prior work has suggested simple captchas [1] for quality control, e.g., "How many paragraphs does this passage contain?" Our pilot study included such a captcha but found it not useful, as later discussed in Section 4.3. Others have also suggested not requiring all questions to be answered and testing for this as a measure of participant effort [41]. However, this seemed unnecessary and inefficient given other controls we already had in place.

Finally, the clarity and accuracy of our questionnaire was pre-tested before the actual pilot test using cognitive interviewing [17]. In particular, we asked ten graduate students (not authors of this study) to read a passage and then answer each question. They were then asked to to re-articulate each question in their own words and explain their answers. Based on this, some problems with initial questions were detected and fixed prior to the pilot. While this comes from survey design, Alsono has also suggested such pre-testing in general for crowdsourcing studies [1].

Our survey was built and hosted on Qualtrics and posted for data collection to Amazon's Mechanical Turk (AMT) platform, with a brief summary of the task and an external link to the survey. We required workers to have a prior 95% approval rate; with best practices suggesting that such approval rating filtering is necessary but not sufficient in and of itself. We did not require any qualification test or exclude any workers by geographic region (many other studies restrict to U.S. participants as a proxy for English language competency or cultural familiarity). Each worker was allowed to complete the survey once for $0.26 payment. While we wanted to bound completion time, there are reports of many AMT requesters setting unreasonably short time limits that anger workers [50]. We thus informed workers that they could also email us with their completion code and worker ID if necessary. Qualtrics provided each respondent with a completion code to be entered into the AMT task form for payment. While prior work has reported seemingly fraudulent resubmission or alternation of codes [24], we did not observe this.

3.2 Search Scenarios & Document Collection

Our study posed three search scenarios for consideration:

- **Health:** *Imagine that you or your friends are trying to make a plan for fast weight loss*
- **Travel:** *Imagine that you will have a holiday in China for seven days*
- **Technology:** *Imagine that you are writing a paper about the influence of smartphones on society.*

Participants self-selected the search scenario to work on, and we pre-selected a set of documents to be judged for each scenario. Each participant was asked to read a randomly selected document and complete the questionnaire for it.

For each scenario, we wrote a short search query and submitted it to a commercial search engine: 1) "methods for rapid weight loss"; 2) "seven day trip to China"; and 3) "impact of smartphones on society". Stratified sampling was then used over Google rankings to approximate decreasing relevance classes [16]. In addition to mitigating relevance bias, stratified sampling was also expected to yield a broad set of documents across relevance dimensions. A Webpage was randomly selected from the top 10% of results, another from the next 10%, and so on, until 10 Webpages had been selected for each scenario. In a real search setting, the actual distribution of relevant vs. non-relevant documents observed by the user could vary greatly, which naturally could influence the user's relevance thresholds (e.g., being more liberal with few relevant results, or more conservative when there are many) [65]. Just as Cranfield assessment assumes documents are judged independently when judging only topicality, we extend this independence assumption to judging multidimensional criteria. Text from each Webpage was extracted and standardized to avoid any visual presentation effects.

3.3 Pilot Study

To pilot our study design, we recruited 86 participants from AMT to complete our survey. As discussed earlier, while our pilot included a captcha as a quality control measure, we found it to be subsumed by the opposite question-pairs controls. In particular, respondents who missed the question-pairs controls typically passed the captcha, suggesting its low utility. Moreover, other prior work has suggested respondents may dislike such captchas [50], and so we discontinued using the captcha after our pilot. We also received worker comments suggesting that the initial time allotted (10 minutes) and payment ($0.05) were insufficient, and therefore we increased the working time to 25 minutes and basic payment to $0.1. We also decided to begin offering a bonus payment of $0.16 for responses passing quality assurance tests.

4. DATA ANALYSIS

Our main study collected surveys from 502 AMT workers. Topics were self-selected by workers (Health: 232, Travel: 93, Technology: 177), possibly reflecting unassessed prior familiarity with the selected topic. We filtered out 118 responses (23.5%)

which failed our quality control tests. Specifically, Likert responses to highly similar questions were required to be +/- 1 of one another. Following prior work, we imposed a stricter criteria for opposite question-pairs, requiring identical responses after scale inversion [9, 27]. Of the remaining 384 responses (Health: 173, Travel: 75, Technology: 136), we partitioned data into two sets: 150 responses were used for Exploratory Factor Analysis (EFA), and 234 participants for the Confirmatory Factor Analysis (CFA). This partition was chosen because 150 and 200 are the minimal recommended observations for applying EFA and CFA, respectively [75]. Many software packages are available for SEM (e.g., `LISREL`, `SPSS` and `SAS`). We used freely-available `R` with the `psych` and `sem` packages (cran.r-project.org/web/packages/).

4.1 Exploratory Factor Analysis (EFA)

Exploratory factor analysis (EFA) is utilized to determine: (a) the number of latent factors underlying responses to the scale items (i.e. survey questions); (b) the specific scale items that measure each factor; (c) the substantive label assigned to each factor; (d) the nature of correlations between the factors [33].

Regarding sample size, accepted practice is to perform Kaiser-Mayer-Olkin (KMO) Measure of Sampling Adequacy and Bartlett's Test of Sphericity to ensure this sample supports valid EFA [75]. KMO "indicates the extent to which a correlation matrix actually contains factors or simply chance correlations between a small subset of variables". Tabachnick and Fidell [66] suggest that values of 0.60 and higher are required. Bartlett's (1950) test of Sphericity is used to estimate the zero correlation probabilities in the matrix. However, this test is very sensitive to sample size and so must also be applied with care[66]. We evaluated both KMO and Bartlett's Test prior to EFA. The result of KMO was 0.870, with Bartlett's Test yielding 2300.44 (p<0.001). Both values indicate that our own sample used satisfies the requisite assumptions for proceeding with EFA [33].

Recall that Section 2.1 hypothesized 5 dimensions of relevance: topicality, novelty, understandability, scope, and reliability. If these dimensions and questions were well matched and crafted, we should observe expected correlations. In particular, questions intended to interrogate a particular dimension should have responses highly correlated with one another, and only weakly correlate with other questions. If so, we would then observe 5 clusters of correlated questions (one per dimension). Each hypothesized dimension would then comprise a distinct *factor* in our ultimate model. In simple terms, EFA enables us to empirically validate or refute these expectations.

The first step in EFA is the initial extraction of the factors to be included in the model. Standard *Maximum likelihood* and *principal axis factoring* (PAF) methods were adopted for extracting factors [53]. PAF is a least-squares estimation of the latent factor model which minimizes the unweighted sum of the squares or ordinary least squares of the residual matrix [74]:

$$F_{\text{OLS}} = \frac{1}{2}\,\text{tr}[(S - \sum)^2] = \sum_j \sum_j (s_{ij} - \sigma_{ij})^2$$

S is the correlation matrix of observed sample. $\sum$ is the model-fitted correlation matrix. s_{ij} are the elements of matrix S, and σ_{ij} the elements of the matrix. Following factors extraction, *rotation* is employed to maximize high correlations between factors and variables and minimize low ones [66]. There are two kinds of rotations in EFA: orthogonal and oblique. Orthogonal rotation assumes no correlations exist between the resulting factors, while oblique rotation allows the factors correlated with others [33]. Given observed correlations, we adopt oblique (Promax) rotation.

Our initial EFA was conducted in *R* assuming 30 latent factors. After applying PAF and oblique rotation, the resulting *Scree plot* for this initial EFA is shown in **Figure 1.** The top 30 Eigenvalues (marked by circles) are computed from the correlation matrix and ordered by decreasing value along the x-axis. The Scree plot of decreasing eigenvalues is used to identify where values roughly level-off [75]. To determine the number of factors to keep, we use *parallel analysis* [25]. A random dataset is generated with the same number of responses and variables as in our sample data. We then created a correlation matrix and computed its eigenvalues. **Figure 1** augments the scree plot with an induced parallel analysis line marked by triangles. The red line shows the best intersection to the Eigenvalue plot based on parallel analysis. We should keep only as many factors as appear *above* it: five.

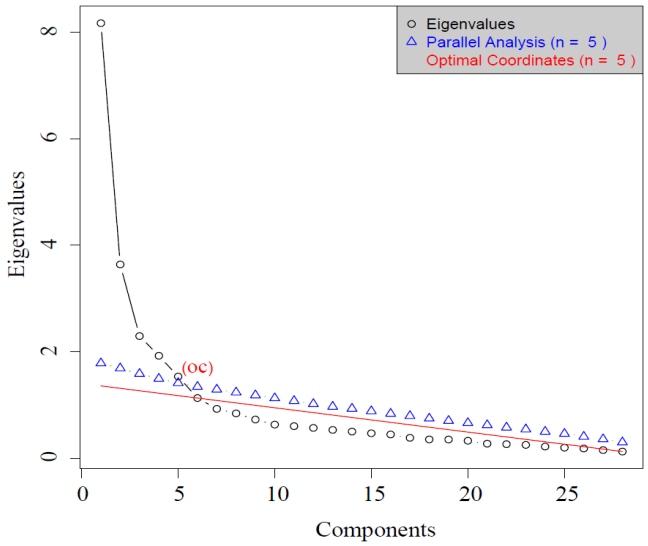

Figure 1. **Revised scree plot showing parallel analysis results**

Given initial EFA supporting inclusion of five factors in the model, EFA was then run again with the number of factors fixed to 5. **Table 1** shows the Pearson correlation r observed between the five factors, ranging in [-0.25, 0.54]. Given standard levels of correlation defined as weak (r=0.1), medium (r=0.4), strong (r=0.7), and very strong (r=0.9) [18], the data is interpreted to show medium correlation among the five factors.

Which questions (i.e., items) should be discarded due to weak correlation, or correlation with multiple factors? Standard EFA practice [75] is to discard questions with: 1) Weak factor loading < 0.4; 2) Cross-loading < 0.15 (difference in estimated loadings across multiple factors); and 3) Lack of logical agreement between question semantics (e.g., a question intended to be about novelty which highly correlated instead with reliability). *Factor loading* regression coefficients quantify direct effects of factors on items. Based on these criteria and *Kenny's Rule* (Section 3.1), we keep three items per factor (e.g., R1-3 load on Reliability).

Table 3 shows the five factors loadings for the items we kept, with *standardized loadings* between [0.52,0.94]. The final column, h^2, denotes the final *communality estimate*: the proportion of variance accounted for by retained factors. A value of h^2 < 0.40 indicates that an item is less strongly correlated with its corresponding factor [75]. The *interpretability criterion* guides us to expect that "the manifest variables appear to cluster together in ways that seem logical and reasonable, given constructs that are being measured" [33]. **Table 3** shows that questions associated with each of the five factors included do cluster as expected.

4.2 Confirmatory Factor Analysis (CFA)

EFA is used in exploratory situations to discover a possible factor structure but not to validate it. To confirm the resulting factors found by EFA, CFA is employed following EFA to assess the goodness of fit between a candidate factor model vs. the actual relationships evidenced in the data [33].

In our model, relevance is represented as a latent factor atop the other relevance dimension factors. A *hierarchical factor model* is proposed, defined according to the following equations:

$$y = \Lambda_y \eta + \varepsilon \qquad \eta = B\eta + \Gamma\xi + \varsigma$$

where Λ_y is the matrix of the loadings for endogenous variables; B is the matrix of causal path; Γ is the matrix of causal path from exogenous to endogenous; ε and ς are the residuals. η represent the exogenous and endogenous latent variables. Maximum likelihood is used to estimate the model parameters as follows:

$$F_{ML} = \log\left|\sum\right| - \log|S| + tr\left(S\sum^{-1}\right) - \rho$$

where ρ is the number of the observed variables; $\sum$ is the estimated covariance matrix of the proposed model and S is the actual covariance matrix of the sample. CFA is a large sample technique. After discarding surveys failing quality tests, 234 remained for analysis. This number exceeded the generally accepted minimum of 200 instances needed [56].

Figure 2 depicts the resulting structure of our proposed *structural equation model* (SEM) of multidimensional relevance. Latent measurement errors (not shown explicitly in the Figure) were assumed, modeled, and estimated in SEM. Observed data (responses to survey questions) are shown in square boxes, with induced factors shown in ovals. Directed edges connect the top-most latent factor, *relevance*, with the five factors selected from EFA (topicality, novelty, understandability, scope, and reliability). These factors are each connected to their respective observed data. Edge weights quantify the inferred *factor loading* relationships.

Table 2 shows the factor loadings resulting from CFA and their statistical significance. *Standardized loadings* shown here match those shown graphically in **Figure 2**, but differ from the factor loadings shown in **Table 3**, since EFA analysis shown there was exploratory to determine model structure, whereas CFA analysis shown here is used to confirm a specific model structure. Standardized loadings range from [-1.00,1.00] and show the strength of correlation. *Unstandardized loadings* determine if standardized loadings are statistically significant from a t-test.

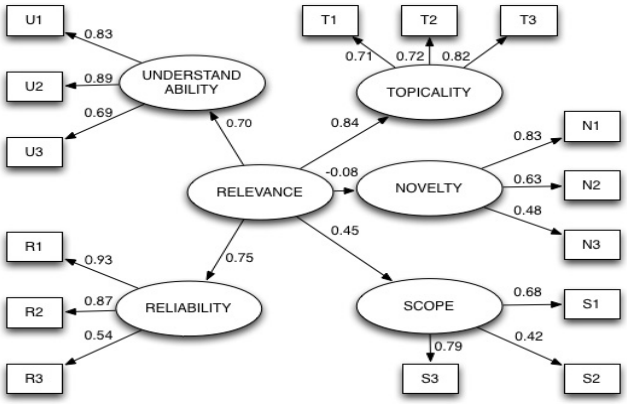

Figure 2. Our structural equation model for modeling relevance.

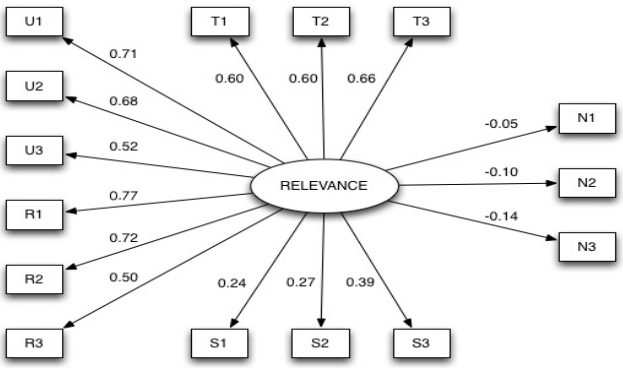

Figure 3. Alternative first-order factor model without factors.

Table 1. Pearson *r* correlation between model factors.

Factor	F_1	F_2	F_3	F_4
F_1 Reliability	--			
F_2 Topicality	.54	--		
F_3 Scope	.46	.34	--	
F_4 Novelty	.19	.19	-.25	--
F_5 Understandability	.27	.26	.26	-.21

Table 2. Factor loadings for each survey question (i.e., item).

Factors & Items	Standard. Loading	Unstd. Loading	t-test: $^*p < .001$
Reliability			
R1	0.93	0.90	10.53*
R2	0.87	0.89	10.57*
R3	0.54	0.63	7.27*
Topicality			
T1	0.71	0.74	6.08*
T2	0.72	0.66	6.13*
T3	0.82	0.80	6.24*
Scope			
S1	0.68	0.92	8.66*
S2	0.42	0.64	5.67*
S3	0.79	1.13	9.12*
Novelty			
N1	0.83	1.45	9.24*
N2	0.63	1.08	7.87*
N3	0.48	0.77	6.39*
Understandability			
U1	0.83	0.74	11.06*
U2	0.89	0.84	11.41*
U3	0.69	0.72	9.48*

We also evaluate two baseline models. The *Null Model* effectively assumes observations are independent, with all covariance between questions fixed to 0 and the means and covariance left free. Our other baseline model, a *first-order factor model* (**Figure 3**), posits no latent factors for relevance dimensions, using a single relevance latent factor to explain observed data. **Table 4** shows global fitness indices for our proposed model vs. the two alternative models. For the proposed model, best χ^2 fit is achieved, and with the fewest degrees of freedom (*df*) as well. The root mean square error of approximation (RMSEA) is 0.065, well below the standard acceptable level of 0.1 [2]. Standardized root mean-square residual (SMSR) is also used to measure fit between model and data. We see our proposed model achieves SMSR of 0.0692, well below the acceptance level of 0.08 [37, 53]. Both Non-Normed Fit index (NNFI) and Comparative Fit Index (CFI)

Table 3. Promax-Rotated Factor Pattern. Factor 1-5 ordering above: reliability, topicality, scope, novelty, and understandability.

Factor						
1	**2**	**3**	**4**	**5**	**h^2**	**Survey Question (Item):** number and text
.94	.01	-.13	-.12	.01	.77	**R1.** I think the information in this passage is wrong.
.82	-.03	-.10	-.02	.16	.67	**R2.** I think some or all of the information in this passage is incorrect.
.68	-.23	.23	-.15	-.29	.48	**R3.** I think the information in this passage needs further proof.
-.09	.88	-.14	.09	.12	.74	**T1.** The topic of this passage relates to the information I am looking for.
-.02	.86	-.02	-0.35	0.08	.68	**T2.** The main content of this passage is not related to the topic I want to know.
-.07	.76	.19	-.15	.00	.65	**T3.** This passage does not mention anything close to the topic I'm interested in.
-.14	-.08	.91	.14	.05	.66	**S1.** The information in this passage is either too general or too specific.
.00	-.10	.71	.05	.02	.45	**S2.** I think I need either more detailed or more generalized information here.
.14	.08	.68	.05	.06	.65	**S3.** The content of this passage is either too broad or too narrow for what I want.
-.06	-.03	.15	.80	.00	.57	**N1.** The information in this passage is very new to me.
-.13	.01	.22	.67	-.04	.39	**N2.** I have heard about such information/ideas/knowledge before.
.03	-.03	-.03	.52	.01	.27	**N3.** This passage is different from others that I have read before.
.02	.07	.06	.13	.80	.69	**U1.** I understand what the author is talking about in this passage.
.13	.00	.07	-.10	.75	.71	**U2.** It's easy for me to understand most of the information in this passage.
-.04	-.05	.16	-.16	.57	.57	**U3.** I'm able to follow the content of this passage with little effort.

Table 4. Global fitness indices of the Proposed model vs. the Null Model and First-order factor model.

Model	**χ^2**	***df***	**χ^2/df**	**NNFI**	**CFI**	**RMSEA**	**SMSR**
Null Model	1429.30	105	13.61	--	--	--	--
First-order FM	671.23	90	7.46	0.486	0.559	0.166	0.121
Proposed Model	168.751	85	1.99	0.922	0.936	0.0650	0.0692

also exceed 0.9. This indicates that > 90% covariance among variables is explained, yielding an acceptable model fit [48].

Figure 2 shows that topicality most strongly impacted relevance, followed by understandability and reliability. Scope weakly contributed, while novelty did not contribute. Consequently, we studied excluding novelty from the model. This improved its fit to observed data: RMSEA was 0.055; Non-Normed Fit index (NNFI) was 0.96 and Comparative Fit Index (CFI) was 0.97. Improvement in χ^2model fit was highly significant (p<<0.001)

5. DISCUSSION

Building upon the excellent framework proposed by Xu and Chen [77] for positivist investigation of multidimensional relevance, we now review key differences between studies and methodology.

To begin, whereas their study asked respondents to directly judge relevance, we do not. As always, there is the question of what is meant by "relevance"? How is the notion to be operationalized, and how do we expect respondents to interpret this term? With explicit judging, guidelines show wide variance: TREC criteria seem lenient in including any document one might cite in a comprehensive report, commercial guidelines are very conservative in restricting which documents satisfy the upper echelons of graded relevance, and untrained judges tend to fall somewhere in the middle [40]. What is not clear is how much we learn about multidimensional relevance at large when relevance is so specifically defined. Should we instead ask for relevance judgments without defining any criteria, permitting wide ranging interpretation? Should we collect judgments for many different operational contexts of relevance decision-making to examine how models change as a function of search task and conditions? As with other forms of data-driven experimentation, we might seek explanatory models which generalize well across several tasks and information needs, yet expect specialized models to show better fit to observed data for specific search scenarios. Xu and Chen define their notion of relevance as "situational relevance" (though their use of this term differs from that of others' [11, 32, 62]). Regardless, this particular notion omits many other important aspects of relevance from prior studies over the decades [51, 62]. We can learn from their study how various factors interact with this particular notion of relevance, but how much can we learn about relevance in general?

Their lab study asked participants to select one of four search scenarios and find as much information as possible through interactive search. After searching, each participant was asked to select two Webpages, each browsed for at least one minute (verified in a user log) and complete a questionnaire for each. As noted earlier, this design yielded skewed data in which more relevant documents were selected. One might ask users to select an equal number of relevant and non-relevant documents, but such selection of non-relevant documents may seem unnatural. While this imbalance impacts how adequately their design handled non-relevant documents, there is a more fundamental methodological issue. With predominantly relevant documents, survey questions would tend to yield skewed Likert responses which violate underlying assumptions of the EFA and CFA analysis used.

While both of our studies had participants self-select the search scenario, our participants were asked to complete the questionnaire for a randomly selected document. Our intent was: 1) to avoid the above bias problem; 2) to be easier to scale via crowdsourcing, 3) to accommodate participant judging

preferences (Sanderson cites Soboroff and Robertson reporting "assessors preferred to assess rather than search" [60]); and 4) to resemble Cranfield-style topical relevance judging [57] which we posit such psychometric methods could usefully support. In particular, we believe psychometric techniques can help generate reliable and large sets of judgments for evaluating search systems, and that multidimensional judgments could inform the long-entrenched issue of judging disagreements [3, 32, 40, 72].

In their study, only 72 document evaluations were used for EFA. This is problematic because covariation patterns could be unstable and might not represent the intended population sufficiently. Kaiser-Mayer-Olkin (KMO) Measure of Sampling Adequacy and Bartlett's Test of Sphericity are typically performed to ensure sufficient sample size. Moreover, while PCA was used to extract factors, PCA is not a valid EFA. While PCA can be used to reduce manifest variables to fewer synthetic variables, it is not appropriate for uncovering the factor structure which underlies a dataset [33]. In addition, despite showing correlations exist between the five factors considered, orthogonal rather than oblique rotation is still used. Finally, the criteria used to determine the factors in their structural equation model are not reported.

For CFA, whereas *path analysis* [77] was used to investigate the relations between the five factors and relevance, path analysis assumes that no measurement error exists in the model. Thus, measurement error was not considered in their model to distinguish observed variables common variance vs. their error variances [33]. Also, six latent factors – relevance and five dimensions of topicality, reliability, understandability, novelty, and scope – are each associated with a set of survey questions. Each latent factor is inferred based on responses to those questions. In the path analysis, however, these six factors are all treated as manifest variables; structurally, causal directed edges point from the five dimensional factors to relevance. Their modeling objective is to predict relevance from the other five factors, essentially a multiple regression task evaluated by R^2 statistics. Finally, their proposed model is not empirically compared to any alternative model (common A/B testing).

In contrast, we focused entirely on relative model fit and not on prediction error; since we did not ask respondents to judge relevance, we could not assess how well our model predicted relevance, an apparent shortcoming. How might we evaluate prediction error? Each option has its own limitations: 1) collect relevance judgments as they did; 2) collect relevance judgments in a separate questionnaire to avoid any interaction between relevance judging vs. other questions; 3) collect relevance judgments from trusted assessors or editors according to some particular relevance criteria; and 4) compare rank correlation between relevance predictions from the model vs. search engine ranking [16] from which documents were sampled (Section 3.2).

With regard to modeling differences, we have already noted our hierarchical vs. their flat modeling of latent factors. Another structural difference is directed causation edges point in the opposite directions: from relevance to the five factors, and from each factor to its corresponding items (see **Figure 2**). On one hand, it seems more intuitive that novelty should (partially) cause relevance (their model), rather than relevance causing novelty in our model. On the other hand, it is more intuitive that in our model that latent novelty should cause the responses to questions about novelty, as opposed to the other way around. Valid inferences may be drawn in both models, and we can let the data speak for itself in regard to how well each model fits the data. That said, further consideration of causality is warranted.

With regard to findings, they showed that novelty and topicality contributed equally to explaining relevance judgments, with lesser contributions from reliability and understandability. Scope did not appear contribute in any meaningful way to relevance. Our own findings, shown in **Figure 2**, are rather different. What might have contributed to such different findings between our two studies?

Their sample of documents were self-selected by participants after interactive-search and biased toward relevance, whereas we assigned documents to be judged and control bias by stratified sampling from search engine results. People have been reported to judge their own search results differently than assigned documents [16]. Their participants were drawn from a university student population and ours from AMT. Our topics were different, intended to be more widely familiar to a distributed crowd. These topical differences may be significant (e.g., novelty would presumably be more important for news-oriented search topics). Prior findings [8, 59, 67, 68, 71] support that people's criteria to relevance judgment change in different tasks, given different cognitive states as a task changes or the stages of a task are switched [62]. Some criteria may dominate in some domains (tasks) while being entirely dispensable in others.

6. CONCLUSION

Understanding the nature of relevance and the various factors that contribute to it is one of the most fundamental and long-standing research questions in information retrieval, yet one in which even today there seems to be little agreement about the number of factors or their relative import. We believe the positivist framework offered by Xu and Chen [77] offers our community an avenue for gaining significant traction, but that some of the mechanics of their originally proposed methodology require refinement to ensure valid conclusions are drawn from studies based on their psychometric approach. While we resolve many of these concerns via revised methodology we have proposed, we have also discussed other questions that remain for future work.

The potential of psychometrics methods for IR extends beyond informing our understanding of factors contributing to end-user relevance judgments. For example, large-scale, multidimensional relevance judging could support more informative evaluation of IR systems beyond traditional Cranfield topicality. Moreover, multidimensional judgment data could yield new understanding of causes leading to disagreement in Cranfield relevance judgments. Prior studies to date have investigated subjective relevance thresholds; varying interpretations of the underlying information need; human factors such as priming, fatigue and boredom; and issues in crowdsourcing like fraud or poor language skills. Complementing this, multidimensional relevance data could: 1) provide new insights into non-topical effects explaining disagreement in supposedly topical judgments; 2) enable more robust inference of relevance judgments as a function of multiple factors; and 3) enable comparative studies of disagreement in topical judging vs. disagreement in judging other factors. Finally, proven quality control methods from psychometrics could enable more robust crowdsourcing data collection in general.

With regard to future work, one might investigate wider relevance factors, search scenarios, users, and topics. We might model negative factors people use in making relevance decisions [29] as well as positive ones, or multi-stage relevance judging [29] rather than assuming a single decision point. Measurement error in survey responses might be estimated by deliberation time or other analytic data available through instrumentation. As prior work has done in crowdsourcing [41], carefully controlled experiments could assess the relative import of different quality control tests or

aggregation strategies, such as aggregating multiple factors at the individual-level rather than one dimensional judgments across individuals. One might then aggregate crowdsourced multi-criteria judgments collected at the individual-level [20].

Cognitively, we still do not know much about how users integrate relevance criteria. Greisdorf's work [29] is the only one we are aware of attempting multi-stage modeling of relevance determinations by users. From the systems-oriented perspective, there is an operational question of how IR systems detect and combine various sources of evidence regarding relevance to induce an overall document ranking. Systems can: 1) infer overall user relevance by observable behaviors like clicks; 2) define and extract features approximating relevance criteria beyond topicality, such as authority or readability; and 3) learn weighting functions for combining features [47]. Eickhoff et al. [23] emphasize non-linear dependencies among such features, and provide pointers to other related work. Tsikrika and Lalmas [70] not only estimate overall relevance from varying relevance criteria, but also infer relative strengths of the different criteria by decomposing overall relevance. The cognitive and systems-oriented IR literatures have been largely disjoint historically, yet there is clearly overlap for further investigation.

Acknowledgments. The anonymous reviewers provided valuable feedback which strengthened this work. This study was supported in part by National Science Foundation grant No. 1253413, DARPA Award N66001-12-1-4256, and IMLS grant RE-04-13-0042-13. Any opinions, findings, and conclusions or recommendations expressed by the authors are entirely their own and do not represent those of the sponsoring agencies.

7. REFERENCES

[1] Alonso, O. 2013. Implementing crowdsourcing-based relevance experimentation: an industrial perspective. *Information Retrieval*. 16, 2, 101–120.

[2] Anderson, J.C. and Gerbing, D.W. 1988. Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*. 103, 3, 411–423.

[3] Bailey, P. et al. 2008. Relevance assessment: are judges exchangeable and does it matter. *SIGIR'08*, 667–674.

[4] Balatsoukas, P. and Ruthven, I. 2012. An eye-tracking approach to the analysis of relevance judgments on the Web: The case of Google search engine. *JASIST*. 63, 9, 1728–1746.

[5] Baokstein, A. 1979. Relevance. *JASIS*. 30, 5, 269–273.

[6] Barry, C.L. 1994. User-defined relevance criteria: An exploratory study. *JASIS*. 45, 3, 149–159.

[7] Barry, C.L. and Schamber, L. 1998. Users' criteria for relevance evaluation: A cross-situational comparison. *IP&M*. 34, 2–3, 219–236.

[8] Bateman, J. 1998. Changes in Relevance Criteria: A Longitudinal Study. *Proceedings of the ASIS Annual Meeting*. 35, 23–32.

[9] Behrend, T.S. et al. 2011. The viability of crowdsourcing for survey research. *Behavior research methods*. 43, 3, 800–813.

[10] Blanco, R. et al. 2011. Repeatable and Reliable Search System Evaluation Using Crowdsourcing. *Proceedings of SIGIR'2011* New York, NY, USA, 923–932.

[11] Borlund, P. 2003. The concept of relevance in IR. *JASIST*. 54, 10, 913–925.

[12] Boyce, B. 1982. Beyond topicality: A two stage view of relevance and the retrieval process. *IP&M* 18, 3, 105–109.

[13] Bradford, S.C. 1934. Sources of information on specific subjects. *Engineering: An Illustrated Weekly Journal (London)*. 137, 26, 85–86.

[14] Browne, M.W. 2000. Psychometrics. *Journal of the American Statistical Association*. 95, 450, 661–665.

[15] Cacioppo, J.T. and Petty, R.E. 1984. The Elaboration Likelihood Model of Persuasion. *Advances in Consumer Research*. 11, 1 673–675.

[16] Chouldechova, A. and Mease, D. 2013. Differences in Search Engine Evaluations Between Query Owners and Non-owners. *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, 103–112.

[17] Cognitive Interviewing: *http://www.uk.sagepub.com/textbooks/Book225856?prodId=Book225856*. Accessed: 2014-01-24.

[18] Cohen, J. 1988. *Statistical power analysis for the behavioral sciences*. L. Erlbaum Associates.

[19] Cool, C. et al. 1993. Characteristics of Texts affecting relevance judgements. *Proceedings of the 14th National Online Meeting*, 77–84.

[20] Da Costa Pereira, C. et al. 2012. Multidimensional relevance: Prioritized aggregation in a personalized Information Retrieval setting. *IP&M*. 48, 2, 340–357.

[21] Cuadra, C.A. and Katter, R.V. 1967. Opening the Black Box of "Relevance." *Journal of Documentation*. 23, 4, 291–303.

[22] Dwyer, J. 2002. *Communication in Business: Strategies and Skills*. Prentice Hall.

[23] Eickhoff, C. et al. 2013. Copulas for Information Retrieval. *Proceedings of SIGIR'2013* (New York, NY, USA), 663–672.

[24] Eickhoff, C. and Vries, A.P. de 2013. Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval*. 16, 2, 121–137.

[25] Franklin, S.B. et al. 1995. Parallel Analysis: a method for determining significant principal components. *Journal of Vegetation Science*. 6, 1, 99–106.

[26] Furr, M. 2011. *Scale Construction and Psychometrics for Social and Personality Psychology*. SAGE.

[27] Goldberg, L.R. and Kilkowski, J.M. 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, 82–98.

[28] Green, R. 1995. Topical relevance relationships. I. Why topic matching fails. *JASIS*. 46, 9, 646–653.

[29] Greisdorf, H. 2003. Relevance thresholds: a multi-stage predictive model of how users evaluate information. *IP&M*, 403–423.

[30] Grice, H.P. 1989. *Studies in the way of words*. Harvard University Press.

[31] Gwizdka, J. 2014. News Stories Relevance Effects on Eye-movements. *Proceedings of the Symposium on Eye Tracking Research and Applications*, 283–286.

[32] Harter, S.P. 1992. Psychological relevance and information science. *JASIS*. 43, 9, 602–615.

[33] Hatcher, L. 2013. *Advanced statistics in research: reading, understanding, and writing up data analysis results*. ShadowFinch Media, LLC.

[34] Hjørland, B. and Christensen, F.S. 2002. Work tasks and socio-cognitive relevance: A specific example. *JASIST*. 53, 11, 960–965.

[35] Hosseini, M. et al. 2012. On Aggregating Labels from Multiple Crowd Workers to Infer Relevance of Documents. *Advances in Information Retrieval*. R. Baeza-Yates et al., eds. Springer Berlin Heidelberg. 182–194.

[36] Hox, J.J. and Bechger, T.M. 2007. An introduction to structural equation modeling.

[37] Hu, L. and Bentler, P.M. 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.

[38] Huang, X. and Soergel, D. 2013. Relevance: An improved framework for explicating the notion. *JASIST*. 64, 1, 18–35.

[39] Johnson, J.R. et al. 1981. Characteristics of Errors in Accounts Receivable and Inventory Audits. *The Accounting Review*. 56, 2, 270–293.

[40] Kazai, G. et al. 2012. An Analysis of Systematic Judging Errors in Information Retrieval. *Proceedings of CIKM'2012* (New York, NY, USA), 105–114.

[41] Kazai, G. et al. 2011. Crowdsourcing for Book Search Evaluation: Impact of Hit Design on Comparative System Ranking. *Proceedings of SIGIR'2011* (New York, NY, USA), 205–214.

[42] Kittur, A. et al. 2008. Crowdsourcing User Studies with Mechanical Turk. *Proceedings of SIGCHI'2008* (New York, NY, USA), 453–456.

[43] Lancaster, F.W. 1968. *Information retrieval systems: characteristics, testing, and evaluation*. Wiley.

[44] Lesk, M.E. and Salton, G. 1968. Relevance assessments and retrieval system evaluation. *Information Storage and Retrieval*. 4, 4, 343–359.

[45] Levitin, A. and Redman, T. 1995. Quality dimensions of a conceptual view. *IP&M*. 31, 1, 81–88.

[46] Little, G. 2009. TurKit: Tools for iterative tasks on mechanical turk. *IEEE Symposium on Visual Languages and Human-Centric Computing,* 252–253.

[47] Liu, T.-Y. 2009. Learning to Rank for Information Retrieval. *Found. Trends Inf. Retr.* 3, 3, 225–331.

[48] M, P. and Bonett, D.G. 1980. Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*. 88, 3, 588–606.

[49] Maron, M.E. 1977. On indexing, retrieval and the meaning of about. *JASIS*. 28, 1, 38–43.

[50] Marshall, C.C. and Shipman, F.M. 2013. Experiences Surveying the Crowd: Reflections on Methods, Participation, and Reliability. *Proceedings of the 5th Annual ACM Web Science Conference*, 234–243.

[51] Mizzaro, S. 1997. Relevance: The whole history. *JASIS*. 48, 9, 810–832.

[52] Moshfeghi, Y. et al. 2013. Understanding Relevance: An fMRI Study. *Advances in Information Retrieval*. P. Serdyukov et al., eds. Springer Berlin Heidelberg. 14–25.

[53] Mueller, R.O. and Hancock, G.R. 2008. Best practices in structural equation modeling. *Best practices in quantitative methods*. 488–508.

[54] Murphy, K.P. 2012. *Machine Learning: A Probabilistic Perspective*. Mit Press.

[55] Pearson - Modern Measurement: Theory, Principles, and Applications of Mental Appraisal, 2/E - Steven J. Osterlind: *http://www.pearsonhighered.com/educator/product/Modern-Measurement-Theory-Principles-and-Applications-of-Mental-Appraisal/9780137010257.page*. Accessed: 2014-01-24.

[56] Principles and Practice of Structural Equation Modeling: Third Edition: *http://www.guilford.com/cgi-bin/cartscript.cgi?page=pr/kline.htm&dir=research/res_quant*. Accessed: 2014-01-24.

[57] Proceedings of the International Conference on Scientific Information -- Two Volumes: *http://books.nap.edu/openbook.php?record_id=10866&page=687*. Accessed: 2014-01-26.

[58] Rees, A.M. and Schultz, D.G. 1967. A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching. *Final Report to the National Science Foundation. Volume* I.

[59] Relevance as process: judgements in the context of scholarly research: *http://www.informationr.net/ir/10-2/paper226*. Accessed: 2014-01-24.

[60] Sanderson, M. 2010. Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval*. 4, 4, 247–375.

[61] Saracevic, T. 2007. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. *JASIST*. 58, 13, 1915–1933.

[62] Saracevic, T. 2007. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *JASIST*. 58, 13, 2126–2144.

[63] Schamber, L. 1994. Relevance and Information Behavior. *Annual Review of Information Science and Technology (ARIST)*. 29, 3–48.

[64] Scheines, R. et al. 1999. Bayesian estimation and testing of structural equation models. *Psychometrika*. 64, 1, 37–52.

[65] Scholer, F. et al. 2013. The Effect of Threshold Priming and Need for Cognition on Relevance Calibration and Assessment. *Proceedings of SIGIR'2013* (New York, NY, USA), 623–632.

[66] Tabachnick, B.G. and Fidell, L.S. 2012. *Using Multivariate Statistics*. Pearson Education, Limited.

[67] Tang, R. and Solomon, P. 1998. Toward an understanding of the dynamics of relevance judgment: An analysis of one person's search behavior. *IP&M*. 34, 2–3, 237–256.

[68] Taylor, A.R. et al. 2007. Relationships between categories of relevance criteria and stage in task completion. *IP&M*. 43, 4, 1071–1084.

[69] The Social Construction of Meaning: An Alternative Perspective on Information Sharing: 2003. *http://pubsonline.informs.org/doi/abs/10.1287/isre.14.1.87.14765*. Accessed: 2014-01-24.

[70] Tsikrika, T. and Lalmas, M. 2007. Combining Evidence for Relevance Criteria: A Framework and Experiments in Web Retrieval. *Advances in Information Retrieval*. G. Amati et al., eds. Springer Berlin Heidelberg. 481–493.

[71] Vakkari, P. and Hakala, N. 2000. Changes in relevance criteria and problem stages in task performance. *Journal of Documentation*. 56, 5, 540–562.

[72] Voorhees, E.M. 1998. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. *Proceedings of SIGIR'1998* (New York, NY, USA), 315–323.

[73] Wilson, D. and Sperber, D. 2002. Relevance Theory. *Handbook of Pragmatics*. G. Ward and L. Horn, eds. Blackwell.

[74] De Winter, J.C.F. and Dodou, D. 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, 695–710.

[75] Worthington, R.L. and Whittaker, T.A. 2006. Scale Development Research A Content Analysis and Recommendations for Best Practices. *The Counseling Psychologist*. 34, 6, 806–838.

[76] Wright, S. *Correlation and causation*.

[77] Xu, Y. (Calvin) and Chen, Z. 2006. Relevance judgment: What do information users consider beyond topicality? *JASIST*. 57, 7, 961–973.

[78] Zuccon, G. et al. 2013. Crowdsourcing interactions: using crowdsourcing for evaluating interactive information retrieval systems. *Information Retrieval*. 16, 2, 267–305.