An Experimental Study on the Effect of Domain Expertise on the Consistency of Relevance Judgements*

주제전문지식이 적합성판정의 일관성에 미치는 영향에 관한 실험적 연구

Stacey Scholten**
Sung-Been Moon (문성빈)***

ABSTRACT

An online experiment was conducted to test the subject-knowledge view of relevance theory in order to find evidence of a conceptual basis for relevance. Six experts in Library and Information Science (LIS), nine Master's students of LIS, and twelve non-experts judged the relevance of 14 abstracts within and outside of the LIS domain. Consistency among the judges was calculated by joint-probability agreement (PA) and interclass correlation coefficients (ICC). When using PA to analyze the judgements, non-experts had a higher consensus regardless of the task or division of groups. However, ICC calculations found Master's candidates had a higher level of consensus than non-experts within LIS, although the experts did not; and the agreement rates on the non-LIS task for all groups were only poor to moderate. It was only when the groups were analyzed as two groups (experts including Master's candidates and non-experts) that the expected trend of higher consistency among experts in the LIS task was seen.

초 록

본 논문은 주제분야 전문지식이 적합성 판단에 미치는 영향을 온라인 실험을 통해 살펴보고 주제분야 전문지식이 적합성개념의 기반이 될 수 있는 지를 검증해 보려고 하였다. 문헌정보학 전문가 6명, 문헌정보학 석사과정 학생 9명, 비전문가 12명이 실험에 참여해 문헌정보학 분야에 대한 14개 논문초록과 문헌정보학 영역 이외 14개 논문초록의 적합성을 판정을 실시하였다. 적합성 판단의 일관성은 공동 확률 일치성(Joint-Probability Agreement, PA)과 IBM SPSS의 클래스간 상관관계 계수(Interclass Correlation Coefficient, ICC)를 통해 산출되었다. PA를 사용한 경우, 비전문가는 과제나 그룹 구분에 상관없이 높은 일관성이 보였다. ICC 계산에 따르면, 문헌정보학 전문가들과 비교하였을 때, 문헌정보학 석사과정학생들은 비전문가들보다 높은 수준의 일관성을 가지고 있다는 것으로 나타났다. 2개 그룹(석사 및 박사를 통합으로 하는 전문가그룹과 비전문가)으로 구분하였을 때는 문헌정보학분야 과제에서 예상대로 전문가들이 더 높은 수준의 일관성을 보이는 경향을 볼 수 있었다.

Keywords: domain expertise, relevance, relevance consistency, subject knowledge view
주제영역전문지식, 적합성, 적합성 일치, 주제지식의 관점

---

* This paper is based on a Master's Thesis of the Dept. of Library & Information Science, Yonsei University.

** Department of Library and Information Science, Yonsei University(scho1318@yonsei.ac.kr) (First author)
*** Professor, Department of Library and Information Science, Yonsei University (sbmoon@yonsei.ac.kr) (Corresponding author)

■ 논문접수일자 : 2021년 8월 16일 ■ 최초심사일자 : 2021년 9월 8일 ■ 게재확정일자 : 2021년 9월 18일
■ 정보관리학회지, 38(3), 1-22, 2021. http://dx.doi.org/10.3743/KOSIM.2021.38.3.001

1. Introduction

The dictionary definition of relevance is “relation to the matter at hand or practical and especially social applicability” (Merriam-Webster, n.d.). Therefore, relevance can be broken down into two parts- relationship and applicability. When conversing with someone, switching topics is not considered good manners unless the two topics are related to one another. The second topic must be relevant to the first. However, even if something is ‘on topic’, it must also apply to the situation at hand, or it is not relevant. In both cases, however, relevance does not need to be clearly defined for people to understand what is relevant to a discussion. People intuitively understand relevance (Saracevic, 1975).

In Information Science (IS), finding relevant information for the user is the ultimate goal of information retrieval (IR) systems. Therefore, relevance is the fundamental concept underlying IR systems (Mizzaro, 1997; Saracevic, 2017). Studies on relevance emerged with the advent of IR systems; and because it was required that relevance be interpreted by a computer, it was usually defined by the coordination level matching of terms. In other words, if a document has a certain percentage of similar terms that are also in the query, it is deemed relevant (van Rijsbergen, 1986). This basic conception of relevance continues to be used. However, throughout the years research has shifted from focusing solely on how to express relevance in an algorithm to be reflected within the system, as in the research conducted by the Text REtrieval Conference (TREC) (Voorhees & Harman, 2005), to attempting to understand the human notion of relevance when using an IR system (Saracevic, 2017).

Despite the relatively long history of research on relevance within IR systems, there remains today a lack of an underlying theory of relevance upon which to base research (Ingerwesen & Jarvelin, 2005; Mizzaro, 1997; Saracevic, 2007; 2017). There have been attempts at frameworks connecting the user experience of relevance to system relevance (Ingerwesen & Jarvelin, 2005; Mizzaro, 1998; van Risjenberg, 1986). However, none of these frameworks have received widespread support (Saracevic, 2017). Varying definitions and frameworks of relevance used in research today make it difficult to come up with an overarching theory of relevance, which in turn leads to a disconnect between research based on different views of relevance. While proving a theory to the extent needed is beyond the scope of this paper, it is hoped this research will be a small stone in building the foundation for relevance by testing a key hypothesis of the subject knowledge view of relevance through an online experiment. The experiment tested Hjørland’s (2002, 268) argument that “the degree of relevance agreement among individuals of a given source should be higher among qualified people in fields in which documents play a well-defined role in connection with human activity based on a well-defined theory.”

2. Literature Review

While user relevance and system relevance both believe that the user’s relevance judgement is the

final word on evaluating relevance, proponents of the subject knowledge relevance, also known as the 'fundamental knowledge view' (Huang & Soergel, 2012) or the 'domain analysis view' (Hjørland, 2010), think differently. Instead, those who support this view believe it is the general consensus of people within a field that should be the determination of relevance. Relevance is intimately connected to the domain structure, and only experts in that domain can accurately determine relevance.

This view can be seen quite early in the literature on relevance. Foskett (1972, 77) stated that there is a difference between 'relevance' and 'pertinence,' where the former can only be "established by the consensus of workers in that field," while the latter is found "in the mind of the user." Unfortunately, this distinction between relevant and pertinent never became prevalent in the literature. Instead, the trends of user relevance research have continued to add more and more variables to the definition of relevance, resulting in Harter's psychological relevance (1992), Saracevic's five manifestations of relevance (1997), Mizzaro's four dimensions of relevance (1998), and many others.

While the subject knowledge view has been overshadowed by the user-view of relevance, there are some current practitioners. Hjørland and Albrechtsen's work (1995) begins by arguing that domain analysis is a latent concept in IS. Hjørland continued upon this line of thought, and in his 2002 work called it a 'socio-cognitive perspective' of IS. Finally, in 2010, he clarified this work and termed it the subject knowledge view of relevance, claiming it was first proposed by Saracevic in his 1975 work. However, if Saracevic ever did hold such a view as Hjørland claims; he no longer does, as evidenced by the fact that in his 2017 book, *The Notion of Relevance in Information Science*, he only introduces two models of relevance: the system view and the user view.

Hjørland, however, has continued to promote this idea of subject knowledge relevance. He states: "The question of relevance is thus primarily to understand the relation between "user needs" and [the] entire information ecology" (Hjørland, 2010, 219). Rather than viewing the user's information need as an inner state, it should instead be seen as simply a lack of knowledge about a specific subject which can be rectified by providing the user with the proper information. Therefore, the user of an IR system does not actually know what is truly relevant to a query and should not be used as the standard for determining relevance.

Both the system view of relevance and the user view of relevance focus on relevance from a narrow standpoint. However, the subject knowledge view of relevance argues that the overall context of both the information (in the system) and user must be taken into account by putting the focus on the topics themselves. Thus, the subject knowledge view argues that if the context of the domain is factored into the problem of relevance, it will give us a clearer understanding of how those individual user factors affect relevance judgements, as well as enable us to more accurately determine how to categorize information sources in such a way that retrieval of relevant documents is improved. In other words, the

subject knowledge view, rather than attempting to overturn either the system view or the user view, seeks to include and underpin both while shifting the viewpoint back to where relevance started- the fundamental idea where relevance is a phenomena that can be known, not just described; and relevance is not between a computer and a document, or a user and a document, but is rather located in the process of communication of information.

Despite the fact that the subject-view of relevance has been around since the 1970s, it has not received much experimental attention or validation. The only explicit use of Hjørland's subject knowledge (domain analysis) framework with regards to relevance in Library and Information Science (LIS) is Van der Veer Martens and Van Fleet's (2012, 944) work which analyzes the norms of relevance work in LIS, and found that "relevance in the aggregate [⋯] is mediated by certain characteristics of the agency and its 'relevance workers.'" However, this was based on content analysis of LIS course syllabi and did not specifically look at IR systems.

Compared to studies explicitly dealing with the subject knowledge view of relevance, there have been a fair amount of studies done on the effect of domain expertise on the information search process, albeit not many focusing on relevance judgements. Most studies about domain expertise (White, Dumais, & Teevan, 2009) focus on the search behavior and simply conclude by saying experts have a more successful or effective search session, without giving their criteria for what makes it a successful search session. While these studies intuitively support the idea of subject knowledge as the basis on which relevance should be founded, there needs to be more research conducted to demonstrate that experts have a higher level of agreement on relevance than non-experts. Luckily, there have been a few studies on the agreement between relevance judgments made by experts and non-experts.

Tamine and Chouquet (2017) examined the differences between medical experts and novices with regards to query formation, relevance judgments, and retrieval performance. They found the perceived task difficulty had an influence on relevance judgments. Novices tended to judge more results as relevant due to their lack of expertise, but experts gave lower relevance scores. In other words, users who had a high level of familiarity with the topic (experts) were able to assess relevance more accurately than those without such familiarity. Finally, the agreement level of relevance was slightly higher among experts, but still low for both groups; however, the conclusion was expertise did play a role in creating consensus amongst the searchers.

Bailey et al. (2008) examined the TREC relevance judgments of three categories of judges: topic originators and task experts, task experts who did not contribute to the topic, and non-expert judges. They determined, unlike most previous research, that disagreement about relevance between judges did have an impact on the retrieval performance of a system and concluded "it is possible that unfamiliarity with task and topic context plays a major role" in disagreement between judges (Bailey et al., 2008, 8).

Dong, Loh, and Mondry (2005) analyzed the effect

of variations in relevance judgements on the performance of a medical retrieval system. They split participants into a gold-standard reviewer (a physician), Group A, which consisted of people trained in Biology or Medicine, and Group B, which was composed of people with no experience, and discovered the common trend where relevance similarity－ the number of evaluators in a group who ranked a document similar to the Gold-standard divided by the total number of evaluators in a group－ was affected by domain knowledge, but the differences did not have a significant impact on the system and its evaluation. They found the difference in relevance similarity between groups was not statistically significant, but it is important to note their relevance judgements were only binary; and they concluded by saying their research should be replicated with scaled relevance judgments.

Finally, Liu and Zhang (2019) investigated what effect a user's prior knowledge had on the quality of their search results in the domain of genetics by asking participants to search on assigned topics and save documents they found to be relevant to such topic. The study determined expert users gave closer relevance scores to the TREC gold-standard judges than the novices. However, it is worth noting their relevance ranking scale was different (5-point) than TREC's ranking scale (3-point), which affects the interpretation of 'closer' relevance scores.

As such, it is evident that, while there have been studies which have incidentally found expertise and relevance judgements are related, there has not been a study which has specifically examined the differences in relevance judgments between experts and non-experts in different domain related tasks. On the other hand, in Saracevic (2017) there were fourteen different studies conducted dealing with relevance consistency over the years. In examining these fourteen studies' methodologies, only Rees and Schultz (1967), Janes (1994), Vakkari and Sormunen (2004), and Ruthven (2014) compared the overlap of relevance judgements made by judges of differing domain knowledge; the rest focused on search experience or did not take specific characteristics into account. According to Saracevic (2017), Rees and Schultz (1967) looked at two different criteria, domain knowledge and search experience, and found that medical experts had the highest rate of consistent relevance judgements on a task related to diabetes. Janes (1994) compared actual users' relevance judgements to three other experimental LIS based groups and found that librarians were able to identify relevance on a level most similar to actual users, followed by experienced LIS students. However, in all cases, non-users ranked the relevance of the documents consistently higher than real-life users. In other words, they overestimated the actual relevance.

Vakkari and Sormunen (2004) compared TREC assessors' judgements to student judgements within the context of examining an interactive retrieval system, but they did not actually examine the relevance assessment consistency between TREC assessors and student judgements. Rather, they noted that in the search expansion process, 46% of documents which had been assessed as non-relevant by TREC assessors were used by the students to expand their query. This demonstrates a sharp difference between expert opinions and non-experts, but no further analysis was done.

Finally, the purpose of Ruthven (2014) was to analyze TREC assessments to better understand the human aspects of relevance. He used three case studies only one of which was on the effect of assessors' characteristics on relevance. He examined the effect of the judge's declared interest, familiarity, knowledge, and confidence about the topic and found the strongest predictor of difference in relevance judgments in TREC was the participant's specific knowledge. The next factor was interest, which was related to familiarity.

While these studies all appear to support Hjørland's claims that domain expertise affects the consistency of relevance judgements, none of them compared the experts and non-experts conducting a domain specific task versus a general task. As such, they cannot prove that the result is not merely a function of similar personality or traits on the part of the user. Despite Saracevic's (2017, 69) claim, "a significant amount of relevance consistency studies were done," upon examination, there has never been a study which deliberately examined the consistency between groups of differing expertise. Consequently, there is a need for experiments related to examining the role of expertise in relevance.

3. Methodology

3.1 Research Framework

The subject knowledge view of relevance as proposed by Hjørland is based on the meta-theory of domain analysis. Domain analysis's main proposition is that the true target of research in information science should be knowledge domains and their structures, not the cognitive processes of individuals (Hjørland & Albrechtsen, 1995). Both systems and users develop together, influencing each other, and are both influenced by the larger theoretical context they are created in (Hjørland, 2010). This allows domain analysis to be applied in multiple ways- within a singular domain, it can be used to look at the underlying theories and structures; and in comparing multiple domains, to examine the varying effects of the structure of the knowledge domain. This is not to say users' individualities should be ignored; nor does it imply the knowledge structure of a domain is fixed and stable (Hjørland, 2010; Hjørland & Albrechtsen, 1995). Rather, knowledge is formed in consensus among a group of individuals who share similar theories about the subject.

What does domain analysis imply for relevance within IR? Hjørland (2010, 229) states, "The approach in the "subject knowledge view" is thus to search for variations in relevance assessments that are connected to basic views or "paradigms" in a domain or across domains." While relevance is influenced by individual factors, it is influenced first and foremost by the larger socio-cognitive context; and while non-experts will judge the relevance of information based on their information needs, the most accurate judges of relevance are experts who know the knowledge structure of their domain.

One other study that used domain analysis as a theory to inform methods for analyzing information behavior is Talja and Maula (2003), which hypothesized a difference in use of e-journals and databases dependent upon

the priority of relevance criteria of the domain. They found primary relevance (topical or paradigmatic) was a more determining factor with regards to the use of e-resources than the domain scatter. Both of this study and Van der Veer Martens and Van Fleet (2012) observed overall trends influenced by domain criteria, in addition to individual differences within the domain. However, neither study focused specifically on examining relevance judgements. Therefore, there still has not been a study explicitly examining how domain knowledge affects relevance judgements.

3.2 Experiment

This study assumes, as previous research has shown, that relevance judgements are affected by domain expertise and as such differences between expert and non-expert groups are to be expected. While expertise in a domain can contain many facets (knowledge, skills, experience), within the confines of this paper knowledge and experience were used as the main criteria for determining the difference between domain experts and non-experts. To determine knowledge, the major of the participants was taken into consideration, and to determine experience, if they were currently working, they were asked whether or not they were working in the LIS field. Therefore, the experiment asked three different groups of participants (non-LIS experts (no knowledge or experience in LIS), LIS Master's candidates and graduates (knowledge but no experience in LIS), and doctoral or Master's graduates currently working in the LIS field (knowledge and experience in LIS)) to categorize the relevance of documents in two separate tasks- one within the field of LIS, and one not. The hypotheses for the experiment are as follows:

- *Hypothesis 1:* Experts in LIS have more consistent relevance judgments on the LIS topic than Master's candidates and graduates of LIS.
- *Hypothesis 2:* Master's candidates and graduates have more consistent judgments on the LIS topic than non-LIS experts.
- *Hypothesis 3:* There is no significant difference between the relevance judgments of experts vs non-experts on the non-LIS task.

3.2.1 Recruitment

The experiment was conducted online from October 8, 2020 to October 20, 2020. LIS participants were recruited through both printed advertisements distributed in Yonsei's LIS labs and in the department office, as well as through SNS channels. Participants were also encouraged to pass on the link to the experiment to other LIS graduate students at different schools. The non-LIS participants were recruited through electronic advertisements in a Yonsei club SNS. In addition, some non-LIS participants were recruited through other LIS participants. All participants were offered the chance to put their name in a drawing for a mobile coupon upon completing the experiment.

Due to the academic nature of the domain, it was determined regular users were not suitable for this task. As such, participation was limited to current undergraduate students and above. In domain analysis, there is a difference between "given" populations, which have been historically or structurally created,

and “constructed” populations, which are based on theory and created by the researcher (Kwon, 2016). This study uses given populations, as determined by the academic major.

3.2.2 Search Task

The difficulty of the search task can have a major effect on relevance judgements. Therefore, while both search tasks were in different domains, they were standardized as Intellectual and Amorphous tasks. Intellectual here refers to the target of the task and is “to enhance the user's understanding of a problem;” while amorphous is the goal of the task, which here is “ill-defined or unclear [and] may evolve along with the user's exploration,” as defined by Jiang (2017, 49), based on the 2012 and 2013 TREC session tracks categorization of search tasks.

In order to make the task consistent among all participants, this study followed the same set-up as Zhitomirsky-Geffet, Bar-Ilan, Levene (2017). The participants were told to assume they were writing a report for two separate classroom assignments, had searched for information on the topics, and were being shown the results of the search. The LIS domain task was defined as creating a report on the factors of user satisfaction of libraries, and the non-LIS domain task was creating a report on the effects of the Korean Wave on other Asian countries.

Rather than asking participants to judge a whole document, abstracts were chosen as a complete but concise representation of domain knowledge (Kwon, 2016). The abstracts were taken from the Yonsei library’s website search engine. For the first task, “도서관 이용자 만족도” (Library User Satisfaction) was searched, and for the second task, “한류 영향” (Korean Wave effects) was searched. Results were selected provided they contained a Korean abstract. The method for selecting the results was the same for both tasks to ensure comparability. The first five hits with a Korean abstract were taken from the first ten results; five more abstracts were taken from result 20 on, and four more from after result 30.

Only fourteen abstracts were presented to prevent any skewing of the relevance results (Huang & Hui-yu, 2004; Purgailis, Parker, & Johnson, 1990). The participants were also informed that the order of the search results had no bearing on the relevance and there was no right or wrong answer for the relevance (Zhitomirsky-Geffet, Bar-Ilan, & Levene, 2017).

3.2.3 Relevance Scale

The participants were presented with a seven-point scale, with one side being labeled “non-relevant” and the other labeled “very relevant.” They were given no definition of what relevance was, but simply asked if the information was relevant to the task given. In research conducted on relevance scaling, anywhere from four to seven-point scales have been found to be acceptable (Tang, Shaw, & Vevea, 1999; Zhitomirsky-Geffet, Bar-Ilan, & Levene, 2018) and both interval and categorical judgments have been found to produce similar results (Spink & Greisdorf, 2001).

3.2.4 Exclusion Criteria

In online experiments, the criteria for determining

which results must be excluded is vital for data integrity (Crump, McDonnell, & Gureckis, 2013). The main exclusion criteria for the experiment was time. Two expert participants were observed while they were taking the experiment to establish a minimum time. Both participants finished in five minutes, so five minutes was established as the minimum time for the whole experiment. Any response that spent less than two minutes per task was excluded. On the other hand, any response that took over 40 minutes for the whole experiment was also excluded because it was assumed the participants were not focused on the task. In addition, any significant time difference between the two tasks was interpreted as the participant being distracted during the task and so was also excluded from data analysis. Finally, giving the same score for each judgement in either task was also pre-chosen as an exclusion criteria; although after exclusion for time criteria no responses of this nature were found.

3.2.5 Data Analysis

This study used two methods for calculating the consensus amongst the relevance judgements of the three groups. First, the traditional joint-probability of agreement was calculated by dividing number of times for each rating assigned by each assessor by the total number of the ratings (Dong, Loh, & Mondry, 2005; Janes, 1994; Rees & Schultz, 1967). Then the two-way mixed, absolute agreement Interclass Correlation Coefficients (ICC) was calculated using IBM SPSS. Within education and clinical medicine, ICCs are commonly used to determine the consistency between raters (Beck et al., 2016; Koo & Li, 2016; Nweke, Perkins, & Afolabi, 2019). As ICC can be used with multiple raters of multiple subjects, and does not exhibit the paradox of Kappa scores (Quarfoot & Levine, 2016), it was chosen as a more reliable method for determining rater consistency.

4. Results

4.1 Participant Demographics

A total of 40 valid responses were collected from the experiment. However, as following the exclusion criteria is vital for ensuring data integrity in online experiments, out of those responses, 13 responses were excluded from the data analysis for violating said criteria (see Table 1). Therefore, a total of 27 responses were analyzed for their inter-rater agreement. The titles of the abstracts, along with each group's average judgment and the standard deviation, can be seen in Appendix A.

As shown in Table 2, the participant demographics were as follows. There were twelve non-experts, nine Master's candidates or graduates in LIS, and six LIS experts. The age range of respondents followed the expected norms. Seventy five percent of the non-experts were between the ages of 18 and 25, with the rest being between 26 and 35 years old; and nine had already graduated from their undergraduate program. There were a variety of majors, including Electrical/Mechanical Engineering, Political Science, and English Language and Literature, but none majored in LIS.

〈Table 1〉 Exclusion Criteria

Exclusion Criteria	Definition	Responses Excluded
Minimum Time	Less than five minutes spent total	9
Task Time	Less than two minutes on either task	1
Maximum Time	Over 35 minutes for experiment	2
Off-task	Large difference between time spent on tasks	1
	Total Excluded	13

〈Table 2〉 Participant Demographics

Non-experts	Age	18-25	26-30	31-35	Total
	N	8	2	2	12

* 9 had graduated from their undergraduate program

Master's candidates and graduates of LIS	Age	26-30	31-35	36-50	Total
	N	5	2	2	9

* 4 had completed their studies but were working in a non-LIS field

Experts	Age	31-35	36-50	51-60	Total
	N	1	3	2	6

* 1 completed Doctoral degree, 2 completed Master's and were working in an LIS field, and 3 were currently in a Doctoral program

Five Master's respondents were between 26 and 30, two between 31 and 35, and two between 36 and 50 years old; four had completed their studies but were working in a non-LIS field. Finally, the LIS experts exhibited the widest range of age, with one between 31 and 35, three between 36 and 50, and two between 51 and 60 years old; only one had completed their Doctoral degree, two had competed their Master's and were working in an LIS field, and three were currently in a Doctoral program.

4.2 Percentage Agreement Analysis

Following traditional methods of determining rater consistency, joint probability agreement (PA) was first calculated. The PA of the three groups was calculated by dividing the number of matches for each rating assigned by each assessor by the total number of the ratings.

Table 3 shows the agreement of each group when judging the relevance of both the LIS abstracts and the abstracts not related to LIS. The highest level of agreement among the groups, for both the LIS task and the non-LIS task, was amongst the non-experts. Hypotheses One and Two, "Experts in LIS will have more consistent relevance judgments on the LIS topic than Master's candidates and graduates of LIS," and "Master's candidates and graduates will have more consistent judgments on the LIS topic than non-LIS experts," were found to be false. While Hypothesis

〈Table 3〉 PA on LIS Task/Non-LIS Task

	LIS Task			Non-LIS task		
	Matches	Total	PA	Matches	Total	PA
Expert	38	210	0.1810	40	210	0.1905
Master's	91	504	0.1806	72	504	0.1429
Non-Expert	204	924	0.2208	184	924	0.1991

〈Table 4〉 Two-group PA on LIS Task/Non-LIS Task

	LIS Task			Non-LIS task		
	Matches	Total	PA	Matches	Total	PA
Expert	292	1470	0.1986	239	1470	0.1626
Non-Expert	204	924	0.2208	184	924	0.1991

Three, "There will be no significant difference between the relevance judgments of experts vs non-experts on the non-LIS task," was found to be true, it is important to note Master's candidates had an appreciably lower level of agreement than both the experts and the non-experts.

Due to the unexpected nature of the findings, it was decided to further analyze the judgements by splitting the participants into two groups, expert and non-expert, under the assumption that Master's candidates and graduates could also be considered as experts in the field of LIS. In other words, the data was reanalyzed by equating expertise with only knowledge, not knowledge and experience. Table 4 shows the results of the PA analysis done on the two groups.

The results of Table 4 show that even when considering both Master's and Doctoral graduates and candidates as experts, the non-experts still show more agreement in their relevance judgements. As such, based on PA, all the hypothesis of the experiment are rejected.

4.3 ICC Analysis

However, one of the arguments against using PA is that it does not take random effects into account. Therefore, the data was also analyzed using IBM SPSS's two-way mixed, absolute agreement ICC. Each group's ICC was calculated separately for both the LIS task and the non-LIS task. Table 5 shows the ICC for the LIS task while Table 6 shows the ICC for the non-LIS task. The traditional interpretation of the ICC score is as follows: values between 0-0.5 have poor reliability; 0.5-0.75 have moderate reliability; values between 0.75-0.9 have good reliability; and values over 0.9 have excellent reliability (Koo & Li, 2016). In addition, the 95% Confidence Interval (CI) shows the 95% chance that the true ICC value lands between the lower and upper bound and is the typical standard for determining the significance of the ICC.

Table 5 displays slightly different results than what was seen with the PA calculations. In this case, the Master's candidates have an ICC score

of 0.769, indicating a good level of reliability, while the experts only have a score of 0.561, which is only moderate reliability. In addition, the expert's 95% CI has a range of 0.674, demonstrating little confidence; whereas the Master's 95% CI range is 0.362, proving more reliable. Hypothesis One was proven false.

However, Table 5 also shows the Master's candidates ICC score of 0.769 is higher than the non-experts ICC score of 0.691, demonstrating a higher level of agreement between their relevance judgements on the LIS task. In the ICC data analysis, the second hypothesis was proven true.

While in the PA calculations, experts and non-experts had similar levels of agreement but the Master's had a significantly lower level, Table 6 groups together Master's and non-experts with a lower level of agreement than the experts. However, for all groups, the 95% CI is wide-ranging, from poor to good agreement. As such, Hypothesis Three was proven true.

In conclusion, when using ICC to calculate the inter-rater agreement, within the LIS task Master's candidates and graduates had the highest level of agreement and experts had the lowest. For the task outside LIS, experts had the highest level of agreement, followed by non-experts. However, it must be noted in all of these cases, the only ICC with a high 95% CI and thus higher significance was the Master's LIS task score.

As when calculating PA, it was decided to further analyze the groups as expert versus non-expert. Tables 7 and 8 show the results for the LIS task and the non-LIS task respectively. As seen in Table 7, when both LIS Master's and PhD respondents are considered together, there is an increase of their ICC to good, whereas the non-experts only have moderate reliability. However, for the non-LIS task, as shown in Table 8, the experts have moderate reliability and the non-experts have poor reliability.

〈Table 5〉 LIS Task ICC

	ICC Average	95% CI	
		Lower	Upper
Expert	.561	0.155	0.829
Master's	.769	0.549	0.911
Non-Expert	.691	0.397	0.882

〈Table 6〉 Non- LIS Task ICC

	ICC Average	95% CI	
		Lower	Upper
Expert	.537	0.134	0.815
Master's	.412	0.019	0.745
Non-Expert	.459	-0.005	0.786

〈Table 7〉 Two-group LIS Task ICC

	ICC Average	95% CI	
		Lower	Upper
Expert	.809	0.635	0.926
Non-Expert	.691	0.397	0.882

〈Table 8〉 Two-group non-LIS Task ICC

	ICC Average	95% CI	
		Lower	Upper
Expert	.598	0.301	0.832
Non-Expert	.459	-0.005	0.786

Table 9 shows the results of the hypotheses in each different data analysis group - percentage agreement between all three groups; percentage agreement between the two adjusted groups; interclass correlation coefficients for all three groups; and interclass correlation coefficients for the two adjusted groups. There is currently research being done on how best to analyze the differences in relevance judgements (Zhitomirsky-Geffet, Bar-Ilan, & Levene, 2017) but this paper chose to use the already established symmetric percentage difference with arithmetic mean when analyzing the PA scores (Cole & Altman, 2017). The percentage difference between the PA scores are shown in the parenthesis. Any difference of 5%+ was chosen as being significant. For instance, the difference between expert and Master's judgements within the LIS field was only 0.22%- a non-significant difference. However, 20% [1] is greater than 5%: the directional alternative hypothesis (H2: Master's candidates and graduates have more consistent judgments on the LIS topic than non-LIS experts) is false, but the opposite direction (-20%) is true. At less than 5%, 4.41% [2] means that the null hypothesis (H3: There is no significant difference between the relevance judgments of experts vs non-experts on the non-LIS task) is true. When compared as two groups, experts and non-experts, the directional alternative hypothesis H2: Master's candidates and graduates have more consistent judgments on the LIS topic than non-LIS experts is still shown to be false (10.6% [3] > 5%), while the opposite direction (-10.6%) is true. Finally, 20.2% [4] is greater than 5%, showing that the null hypothesis (H3: There is no significant difference between the relevance judgments of experts vs non-experts on the non-LIS task) is false.

For the ICC scores in Table 9, an asterisk (*) denotes where there was a difference in the 95% CI reliability rating. The difference in the 95% CI reliability rating was determined to be an acceptable measure of significance as it is the criteria for determining the reliability of the measure, and the less reliable the measure, the less significant the difference. For example, within the LIS task, the expert's 95% CI scores ranged from poor reliability to good reliability, but Master's was

〈Table 9〉 Hypothesis Results

	PA (Expert, Master's, Non-Expert)	PA (Expert, Non-Expert)	ICC (Expert, Master's, Non-Expert)	ICC (Expert, Non-Expert)
H1	False (0.22%)	NA	False*	NA
H2	False (20%)(-)[1]	False (10.6%)(-)[3]	True*	True*
H3	True (4.41%)(-)[2]	False (20.2%)[4]	True	True

from moderate to excellent. Therefore, H1 and H2 are marked with an asterisk, showing that H1 and H2 were selected because there is difference between the groups' ratings. However, both the expert and the non-expert 95% CI scores ranged from poor reliability to good reliability when judging non-LIS abstracts. In other words, while their ICC average itself was different, the range of possible scores was not, and as such the ICC average difference is not significant. Therefore, H3 is not marked with an asterisk.

In summary, as shown in Table 9, when assuming a difference between LIS PhD and Master's students (taking experience into account), not all of the hypotheses for the experiment were proven true; when assuming all graduate students of LIS have a similar level of expertise, the trend hypothesized by Hjørland (2002) is seen when using ICC analysis, although not when using PA. As such, based on this experiment, it is difficult to conclude that experts have a higher consistency in their relevance judgements when judging a task within their field, but do not show such consistency on a task outside their field.

5. Discussion and Conclusion

This research sought to explore the relationship between relevance and domain knowledge by testing a key hypothesis of the subject knowledge view. However, the experiment results were inconclusive. When using PA to analyze the data, non-experts consistently had higher consensus regardless of the task or division of groups, and only once (H3) was a hypothesis selected as "True". On the other hand, using the more widely-used ICC found that Master's candidates and graduates of LIS had a reliably higher level of agreement than non-experts on the LIS task, although the experts did not, thereby rejecting H1; and the agreement rates on the non-LIS task, while proving H3, did not provide further clarification. Nevertheless, with ICC, upon further analysis based upon a change in the definition of expert to only considering knowledge, not both knowledge and experience, the expected trend of higher consistency for experts in the related field task and little difference in the non-field task was seen, with both H2 and H3 being selected.

How should this be interpreted? First, due to the academic nature of abstracts, it is possible the format of the information had an effect on how participants judged relevance. It is presumed undergraduates are not exposed to abstracts as much as graduate students, and as such undergraduates might have used heuristics to determine the relevance, thereby leading

to higher consistency in their judgements. Second, Hjørland's (2002, 268) hypothesis explicitly states it is limited to "fields in which documents play a well-defined role in connection with human activity based on a well-defined theory." While the physical sciences often meet such criteria, the domains of the social sciences are not as well defined. However, this experiment was a contribution to the literature in attempting to go beyond the normal domains typically studied in research relevance.

As with all studies, this paper has limitations. The small nature of the LIS field, combined with the mentally intensive task of making relevance judgements, meant experts in LIS were difficult to recruit. While the experiment results based on the LIS domain were debatable, it is important research continues to examine the difference in relevance agreement between experts in a domain and non-experts. Further research should be completed with larger sample sizes and in different subjects. There has been a considerable lack of research on topics beyond medical science and health, and further research needs to be done on not only the physical sciences, but in the social sciences as well.

References

Bailey, P., Craswell, N., Soboroff, I., Thomas, P., Vries, A.D., & Yilmaz, E. (2008). Relevance assessment: are judges exchangeable and does it matter. Proceedings of the 31st annual International ACM Sigir Conference on Research and Development in Information Retrieval, 667-674. http://doi.org/10.1145/1390334.1390447

Beck, S., Ruhnke, B., Issleib, M., Daubmann, A., Harendza, S., & Zöllner, C. (2016). Analyses of inter-rater reliability between professionals, medical students and trained school children as assessors of basic life support skills. BMC Medical Education, 16(1), 263. http://doi.org/10.1186/s12909-016-0788-9

Cole, T. J. & Altman, D. G. (2017). Statistics notes: What is a percentage difference? BMJ: British Medical Journal (Online), 358. http://doi.org/10.1136/bmj.j3663

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating amazon's mechanical turk as a tool for experimental behavioral research. PloS One, 8(3), e57410-e57410. http://doi.org/10.1371/journal.pone.0057410

Dong, P., Loh, M., & Mondry, A. (2005). Relevance similarity: An alternative means to monitor information retrieval systems. Biomedical Digital Libraries, 2(1), 6-6. http://doi.org/10.1186/1742-5581-2-6

Foskett, D. (1972). A note on the concept of "relevance". Inform. Star. Retr., 8, 77-78. http://doi.org/10.1016/0020-0271(72)90009-5

Harter, S. (1992). Psychological Relevance and Information Science. Journal of the American Society for

Information Science, 43(9), 602-615. http://doi.org/10.1002/(SICI)1097-4571(199210)43:9<602::AID-ASI3>3.0.CO;2-Q

Hjørland, B. & Albrechtsen, H. (1995). Toward a new horizon in information-science - domain-analysis. Journal of the American Society for Information Science, 46(6), 400-425. http://doi.org/10.1002/(SICI)1097-4571(199507)46:6<400::AID-ASI2>3.0.CO;2-Y

Hjørland, B. (2002). Epistemology and the socio-cognitive perspective in information science. Journal of the American Society for Information Science and Technology, 53(4), 257-270. http://doi.org/10.1002/asi.10042

Hjørland, B. (2010). The foundation of the concept of relevance. Journal of the American Society for Information Science and Technology, 61(2), 217-237. http://doi.org/10.1002/asi.21261

Huang, M. & Hui-yu, W. (2004). The influence of document presentation order and number of documents judged on users' judgments of relevance. Journal of the American Society for Information Science and Technology, 55(11), 970-979. http://doi.org/10.1002/asi.20047

Huang, X. & Soergel, D. (2012). Relevance: An improved framework for explicating the notion. Journal of the American Society for Information Science and Technology, 64(1), 18-35. doi:10.1002/asi.22811

Ingerwesen, P. & Jarvelin, H. (2005) Information retrieval in context: IRiX. ACR SIGIR Forum, 39(2), 31-39. http://doi.org/10.1145/1113343.1113351

Janes, J. W. (1994). Other people's judgments: A comparison of users' and others' judgments of document relevance, topicality, and utility. Journal of the American Society for Information Science (1986-1998), 45(3), 160. http://doi.org/10.1002/(SICI)1097-4571(199404)45:3<160::AID-ASI6>3.0.CO;2-4

Jiang, J. (2017). Ephemeral Relevance and User Activities in a Search Session. Doctoral dissertation, University of Pittsburg, United States. Available: http://d-scholarship.pitt.edu/30612/

Koo, T. K. & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155-163. http://doi.org10.1016/j.jcm.2016.02.012

Kwon, H. (2016). On the social epistemological nature of questions: A comparison of knowledge domains' question formulations on the topic of “memory”. Doctoral dissertation, Rutgers, United States. http://doi.org/doi:10.7282/T36Q20DB

Liu, J. & Zhang, X. (2019). The role of domain knowledge in document selection from search results. Journal of the Association for Information Science and Technology, 70(11), 1236-1247. http://doi.org/10.1002/asi.24199

Merriam-Webster. (n.d.). Relevance. In Merriam-Webster.com Dictionary. Available: https://www.merriam-webster.com/dictionary/relevance

Mizzaro, S. (1997). Relevance: The whole history. Journal of the American Society for Information Science, 48(9), 810-832. http://doi.org/10.1002/(SICI)1097-4571(199709)48:9<810::AID-ASI6>3.0.CO;2-U

Mizzaro, S. (1998). How many relevances in information retrieval?. Interacting with Computers, 10(3), 303-320. http://doi.org/10.1016/S0953-5438(98)00012-5

Nweke, W. C., Perkins, T. P., & Afolabi, C. Y. (2019). Reliability analysis of complementary assessment tools for measuring teacher candidate dispositions. Georgia Educational Researcher, 16(2), Article 2. http://doi.org/10.20429/ger.2019.160202

Quarfoot, D. & Levine, R. (2016). How robust are multirater interrater reliability indices to changes in frequency distribution?. The American Statistician, 70(4), 373-384. http://doi.org/10.1080/00031305.2016.1141708

Rees, A. & 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 II, Appendices. Springfield, VA: Clearinghouse for Federal Scientific and Technical Information.

Ruthven, I. (2014). Relevance behaviour in TREC. Journal of Documentation, 70(6), 1098-1117. http://doi.org/10.1108/JD-02-2014-0031

Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. Journal of the American Society for Information Science, 26(6), 321-343. http://doi.org/10.1002/asi.4630260604

Saracevic, T. (1997). The stratified model of information retrieval interaction: Extension and applications. Proceedings of the American Society for Information Science, 34, 313-327. Available: https://www.researchgate.net/publication/333293923_The_stratified_model_of_information_retrieval_interaction_Extension_and_applications

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. Journal of the American Society for Information Science and Technology, 58(13), 1915-1933. http://doi.org/10.1002/asi.20682

Saracevic, T. (2017). The Notion of Relevance in Information Science: Everybody knows what relevance is. But, what is it really? San Rafael, CA: Morgan and Claypool Publishers.

Spink, A. & Greisdorf. H. (2001). Regions and levels: measuring and mapping users' relevance judgments. Journal of the American Society for Information Science and Technology, 52(2), 161-173. http://doi.org/10.1002/1097-4571(2000)9999:99993.0.CO;2-L

Talja, S. & Maula, H. (2003). Reasons for the use and non-use of electronic journals and databases: A domain analytic study in four scholarly disciplines. Journal of Documentation, 59(6), 673. http://doi.org/10.1108/00220410310506312

Tamine, L. & Chouquet, C. (2017). On the impact of domain expertise on query formulation, relevance assessment and retrieval performance in clinical settings. Information Processing and Management, 53(2), 332-350. http://doi.org/10.1016/j.ipm.2016.11.004

Tang, R., Shaw, W. M., & Vevea, J. L. (1999). Towards the identification of the optimal number of relevance categories. Journal of the American Society for Information Science, 50(3), 254-264. http://doi.org/10.1002/(SICI)1097-4571(1999)50:3<254::AID-ASI8>3.0.CO;2-Y

Vakkari, P. & Sormunen, E. (2004). The influence of relevance levels on the effectiveness of interactive information retrieval. Journal of the American Society for Information Science and Technology, 55(11), 963-969. http://doi.org10.1002/asi.20046

Van der Veer Martens, B. & Van Fleet, C. (2012). Opening the black box of “relevance work”: A domain analysis. Journal of the American Society for Information Science and Technology, 63(5), 936-947. http://doi/org10.1002/asi.21699

Van Rijsbergen, C. J. (1986). A new theoretical framework for information retrieval. ACM SIGIR Forum, 21(1-2), 23-29. http://doi.org/10.1145/3130348.3130354

Voorhees, E. M. & Harman, D. K. (Eds.). (2005). TREC: Experiment and evaluation in information retrieval. Cambridge, MA: MIT Press.

White, R., Dumais, S., & Teevan, J. (2009). Characterizing the influence of domain expertise on web search behavior. Proceedings of the Second ACM International Conference on Web Search and Data Mining, 132-141. http://doi.org/10.1145/1498759.1498819

Zhitomirsky-Geffet, M., Bar-Ilan, J., & Levene, M. (2017). Analysis of change in users' assessment of search results over time. Journal of the Association for Information Science and Technology, 68(5), 1137-1148. http://doi.org/10.1002/asi.23745

Zhitomirsky-Geffet, M., Bar-Ilan, J., & Levene, M. (2018). Categorical relevance judgment. Journal of the Association for Information Science and Technology, 69(9), 1084-1094. http://doi.org/10.1002/asi.24035

Appendix A.
Library User Satisfaction

1) 도서관 서비스 품질 평가를 통한 전문도서관 이용자 만족도 연구

	Experts	Master's	Non-experts
Standard Deviation	0.90	1.05	1.21
Average Judgement	6.17	6.00	4.83

2) 지식: 도서관 블로그 서비스의 이용자 만족도 연구

	Experts	Master's	Non-experts
Standard Deviation	0.94	1.69	1.28
Average Judgement	3.67	4.78	4.83

3) 학교도서관 이용자 만족도 조사에 대한 학교도서관 전문인력의 인식에 대한 연구

	Experts	Master's	Non-experts
Standard Deviation	1.70	1.34	1.37
Average Judgement	3.33	3.56	3.67

4) 도서관 이용자 만족도를 매개변수로 하는 이용자 충성도에 관한 연구: K대학 사례

	Experts	Master's	Non-experts
Standard Deviation	1.37	1.10	1.29
Average Judgement	4.33	4.11	5.00

5) 어린이도서관 웹사이트 이용자 만족도 분석: D 어린이도서관 사례중심으로

	Experts	Master's	Non-experts
Standard Deviation	1.11	1.83	1.63
Average Judgement	5.33	4.44	5.00

6) 인적서비스 이용자 만족도 및 지속의도의 이해: 대학도서관의 연구

	Experts	Master's	Non-experts
Standard Deviation	1.37	1.20	1.49
Average Judgement	5.67	5.89	4.67

7) 서비스품질지각에 기반한 대학도서관 이용자 만족도와 충성도 분석

	Experts	Master's	Non-experts
Standard Deviation	1.07	1.57	0.75
Average Judgement	5.83	5.56	5.67

8) 대학도서관의 이용자만족도와 충성도에 관한 연구

	Experts	Master's	Non-experts
Standard Deviation	1.26	1.29	0.99
Average Judgement	5.50	5.89	6.17

9) 제주지역 공공도서관 문화프로그램 실태분석 및 이용자 만족도 연구

	Experts	Master's	Non-experts
Standard Deviation	1.53	1.29	1.42
Average Judgement	5.00	3.11	4.25

10) 공공도서관의 운영방식 및 위탁방식에 따른 이용자 만족도 비교

	Experts	Master's	Non-experts
Standard Deviation	1.86	1.69	1.44
Average Judgement	4.83	4.78	4.08

11) 공공도서관의 이용자만족도에 관한 한.미간 비교사례연구

	Experts	Master's	Non-experts
Standard Deviation	1.77	1.13	1.50
Average Judgement	5.17	5.78	4.92

12) 우리나라 공공도서관의 이용자만족도에 관한 연구: 2010 공공도서관 운영 평가 이용자만족도 조사 결과를 중심으로

	Experts	Master's	Non-experts
Standard Deviation	0.75	1.69	0.82
Average Judgement	6.33	5.78	6.00

13) 대학도서관 OPAC2.0 서비스 이용자 만족도와 중요도에 관한 연구: A와 B대학도서관 도서검색결과를 중심으로

	Experts	Master's	Non-experts
Standard Deviation	0.82	1.47	1.42
Average Judgement	5.00	3.22	4.75

14) 대학도서관 이용자의 공동체의식이 이용자 만족도 및 충성도에 미치는 영향 연구

	Experts	Master's	Non-experts
Standard Deviation	1.57	1.66	0.92
Average Judgement	5.17	5.11	4.75

Korean Wave Effect

1) 중국인의 한국음식 인지도가 한식구매의도에 미치는 영향: 한류 조절변수를 중심으로

	Experts	Master's	Non-experts
Standard Deviation	0.69	1.94	1.79
Average Judgement	6.17	4.67	4.25

2) 중국 예능 방송의 한류 영향 분석 연구

	Experts	Master's	Non-experts
Standard Deviation	0.47	1.34	1.95
Average Judgement	6.67	4.44	4.83

3) 드라마의 주인공 배우, 브랜드, 소비자의 이상자아가 브랜드 구매의도에 미치는 영향: 한류 드라마를 시청하는 중국 소비자를 중심으로

	Experts	Master's	Non-experts
Standard Deviation	1.00	1.29	1.75
Average Judgement	6.00	4.89	4.58

4) 중국 시나 웨이보에서의 한국엔터테인먼트 정보 이용이 한류 콘텐츠 및 한류 호감도에 미치는 영향

	Experts	Master's	Non-experts
Standard Deviation	1.38	1.49	1.82
Average Judgement	4.50	3.67	4.83

5) 반(反)한류 정책이 중국 내 한류에 미치는 영향

	Experts	Master's	Non-experts
Standard Deviation	1.91	0.99	1.78
Average Judgement	5.00	2.89	4.00

6) 중국 화장품 광고에서 한류 텍스트와 비주얼 메시지가 구매의도에 미치는 영향

	Experts	Master's	Non-experts
Standard Deviation	1.26	1.63	1.25
Average Judgement	5.50	4.33	5.33

7) 소비자의 브랜드 신뢰, 가치의식, 원산지 중요성 및 한류 제품에 대한 태도가 구매의도에 미치는 영향: 베트남 소비자 연구

	Experts	Master's	Non-experts
Standard Deviation	0.94	1.76	1.75
Average Judgement	5.67	4.67	4.58

8) 한류 이미지가 한국교육상품의 구매의도에 미치는 영향

	Experts	Master's	Non-experts
Standard Deviation	0.94	1.29	0.58
Average Judgement	5.67	4.89	6.00

9) 신(新) 한류 문화콘텐츠와 경제한류가 한국에 대한 이미지와 태도 및 방문의도에 미치는 영향

	Experts	Master's	Non-experts
Standard Deviation	1.07	1.20	1.61
Average Judgement	5.17	4.89	5.42

10) 심리적 거리가 한류 선호도와 한국 국가이미지에 미치는 영향

	Experts	Master's	Non-experts
Standard Deviation	1.49	1.47	1.65
Average Judgement	3.67	3.78	3.67

11) 한류콘텐츠 이용정도가 중국인의 혐한정서에 미치는 영향: 한류호감 한국인에 대한 긍정적 인식의 이차매개효과검증을 중심으로

	Experts	Master's	Non-experts
Standard Deviation	1.41	1.71	1.04
Average Judgement	6.00	4.44	5.50

12) 아르헨티나의 한류 인식과 경험이 한국 패션 제품의 태도에 미치는 영향

	Experts	Master's	Non-experts
Standard Deviation	1.15	2.18	2.24
Average Judgement	5.00	3.89	4.00

13) 국가 간 거리가 K-Pop 한류 콘텐츠의 온라인 글로벌 확산에 미치는 영향: '강남스타일'을 중심으로

	Experts	Master's	Non-experts
Standard Deviation	1.29	1.71	1.55
Average Judgement	5.00	4.44	4.58

14) 진지한 여가로서의 한류 콘텐츠 소비, 한국 관광지 이미지 및 관광 의도의 영향 관계: 필리핀인 잠재 관광객 사례

	Experts	Master's	Non-experts
Standard Deviation	1.63	1.55	1.75
Average Judgement	5.00	3.22	4.67